The Environmental Consequences of Generative AI LLMs

Key ideas

- Generative AI has had a drastic growth in the recent years
- Chatbots have taken the world by storm
- These models require data and resources to train and maintain
- Has its recent growth been sustainable?

```
  Welcome to
                 EEEEEE  LL      IIII   ZZZZZZ   AAAAA
                 EE      LL       II       ZZ   AA   AA
                 EEEEE   LL       II     ZZZ    AAAAAAA
                 EE      LL       II    ZZ      AA   AA
                 EEEEEE  LLLLLL  IIII ZZZZZZ    AA   AA

  Eliza is a mock Rogerian psychotherapist.
  The original program was described by Joseph Weizenbaum in 1966.
  This implementation by Norbert Landsteiner 2005.


ELIZA: Is something troubling you ?
YOU:   Men are all alike.
ELIZA: What is the connection, do you suppose ?
YOU:   They're always bugging us about something or other.
ELIZA: Can you think of a specific example ?
YOU:   Well, my boyfriend made me come here.
ELIZA: Is it important to you that your boyfriend made you come here ?
YOU:   He says I'm depressed much of the time.
ELIZA: I am sorry to hear that you are depressed.
YOU:   It's true. I am unhappy.
ELIZA: Can you explain what made you unhappy ?
```

Literature

- An MIT paper giving insights into aspects of Generative AI [1]
- GPT-4 Technical Report [2]
- Experiments on watts generated by inferences [3]
- Regulations being put in place for Gen AI [4]
- US datacenter resource reports [5]

Research gap

- Industrial focus for the past 2 years
- Little research into its environmental costs and sustainability
  - Is development as expensive as companies make it to be?
- Data centre data on Gen AI is limited
- Legislations covers AI in general

Future research

Build a chatbot model to analyse resource use

- A lot of estimates on current models
- Model with transparent specifications
- Trial with live users
- Scale to a corporate size

Questions?

References used

1. N. Bashir *et al.*, "The climate and sustainability implications of generative AI," *An MIT Exploration of Generative AI* , 2024. doi: 10.21428/e4baedd9.9070dfe7. [Accessed: Feb. 28, 2025].
2. OpenAI, "GPT-4 technical report," OpenAI, Mar. 2023. [Online]. Available: https://cdn.openai.com/papers/gpt-4.pdf. [Accessed: Feb. 28, 2025].
3. A. S. Luccioni et al., "*Power hungry processing: Watts driving the cost of AI deployment?*" arXiv preprint, arXiv:2311.16863, 2023. [Online]. Available: https://arxiv.org/pdf/2311.16863. [Accessed: Feb. 28, 2025].
4. McKinsey & Company, "As Gen AI advances, regulators and risk functions rush to keep pace," McKinsey & Company, Feb. 2024. [Online]. Available: https://www.mckinsey.com/capabilities/risk-and-resilience/our-insights/as-gen-ai-advances-regulators-and-risk-functions-rush-to-keep-pace. [Accessed: Feb. 28, 2025].
5. A. Shehabi *et al.*, "2024 United States Data Center Energy Usage Report," eScholarship, University of California, 2023. [Online]. Available: https://escholarship.org/uc/item/32d6m0d1. [Accessed: Feb. 28, 2025].