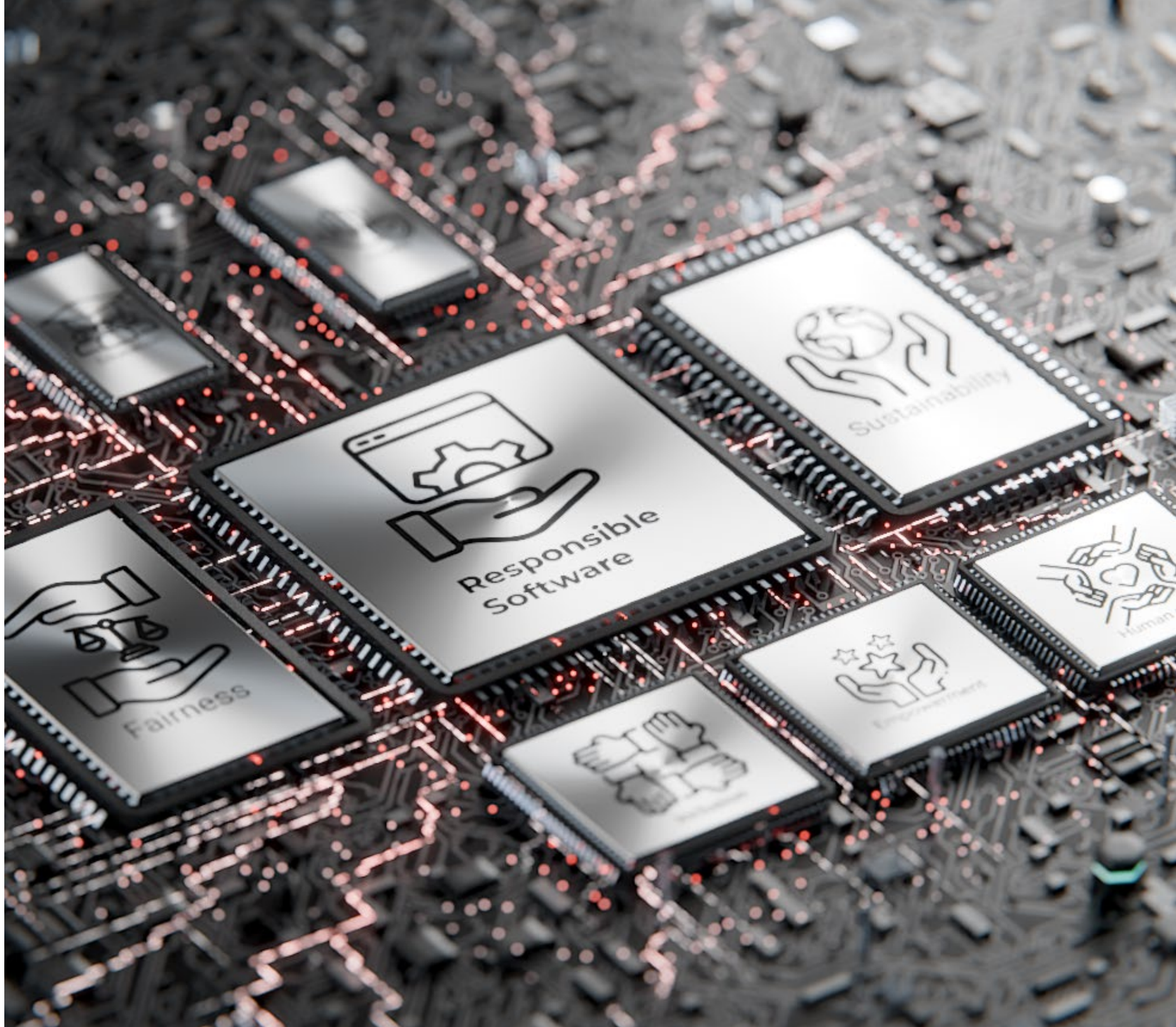
EPFL

Graded 2 Debriefing

2 dec.

Cécile Hardebolle

Responsible Software



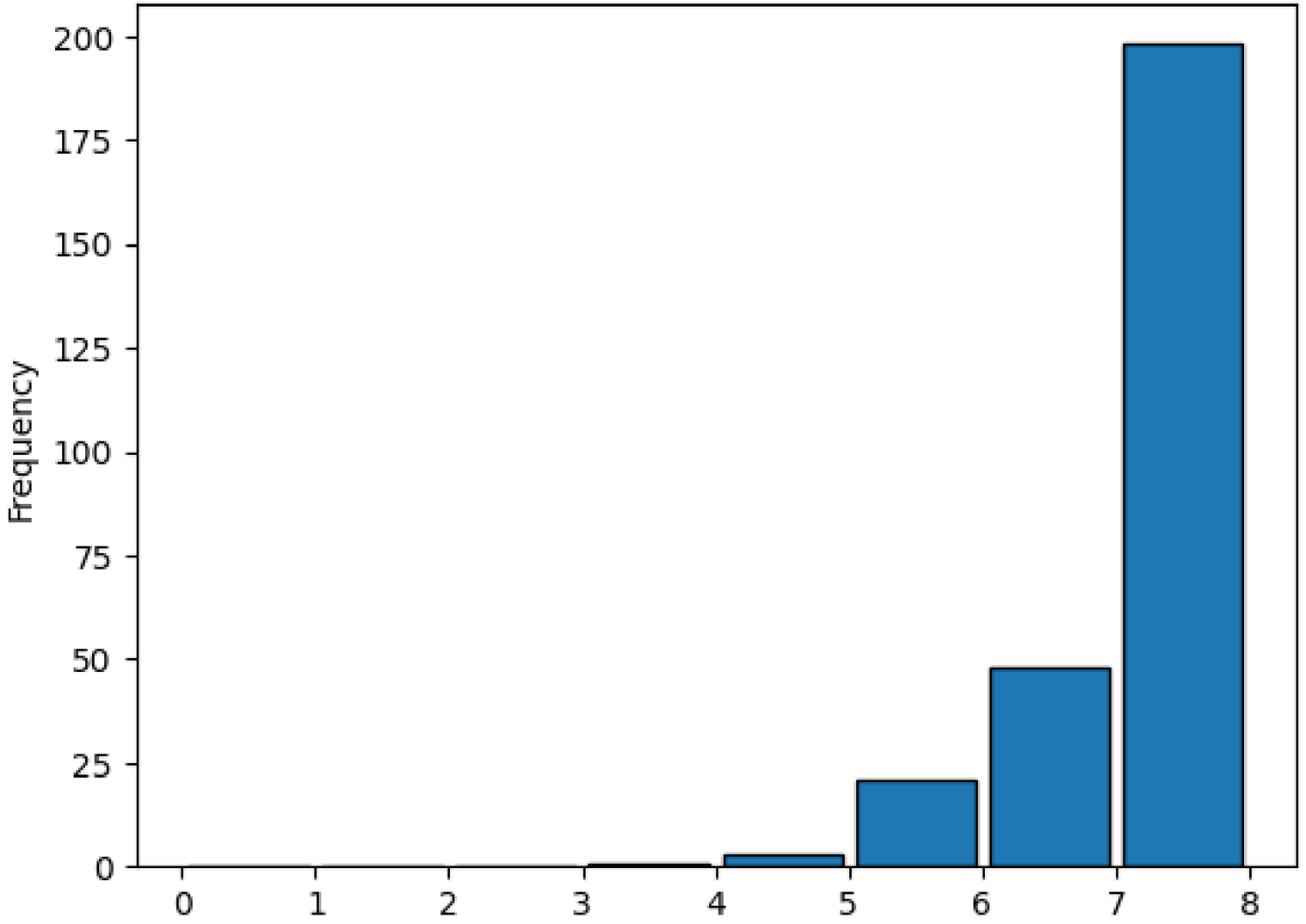
Agenda for today

1. Feedback on the Graded 2 assignment

2. Next dates:
   Empowerment 2 + Conclusion + Revisions

Feedback on Graded 2

Programming questions



Maximum possible:
8 points

Mean: 7.2 points
Median: 7.4 points
(std: 0.8 points)

Exercise 1

Chatbot Model Sustainability Analysis

Questions which created more difficulty

- 1.2.1 (code) Problem variables
- 1.2.2 (code) Training emissions
- 1.2.3 (code) Inference emissions
- 1.2.4 (code) Total emissions

- 1.3.1 (open) Model comparison
- **1.3.2 (open) Ethical values & dilemma**

- 1.4.1 (open) Influence of number of users

1.2.1 Problem variables

Assume that the GPU used both for training and inference consumes 450 W, and the carbon emissions from electricity represent 260 g CO2e/kWh. 1 user makes 5 requests per day with an average of 1000 tokens per request. For this computation, we assume the model is used by 10 000 users per day during 365 days.

gpu_power = 0.450 (in **kW**)
average_token_per_request = 1000
nb_request_per_user_per_day = 5
nb_user_per_day = 10000
total_duration = 365
carbon_intensity = 0.260 (in **kg CO2e / kWh**)

1.2.2 Training emissions

	Model	Training Time (GPU hours)	Model Size (Parameters, in B)	Output Speed (tokens/sec)	Accuracy (%)	Number GPUs Required for Inference
0	Model 1	1000000	175	120	85.2	8
1	Model 2	400000	30	72	78.4	3
2	Model 3	100000	6	120	72.1	1
3	Model 4	570000	70	180	81.0	4

$$\text{Total electricity training} = \text{total number of GPU hours required for training} \times \text{single GPU power}$$

- Access to column “Training Time (GPU hours)”
- Power of **1** GPU

+ Multiply by carbon intensity

1.2.3 Inference emissions

	Model	Training Time (GPU hours)	Model Size (Parameters, in B)	Output Speed (tokens/sec)	Accuracy (%)	Number GPUs Required for Inference
0	Model 1	1000000	175	120	85.2	8
1	Model 2	400000	30	72	78.4	3
2	Model 3	100000	6	120	72.1	1
3	Model 4	570000	70	180	81.0	4

$$\text{Total electricity inference} = \underbrace{\frac{\text{Total number of tokens generated over 1 year}}{\text{Speed of the model in tokens per second}}}_{\text{Total computation time in seconds}} \times \underbrace{\frac{1}{3600}}_{\text{Conversion in hours}} \times \underbrace{\text{Number of GPUs} \times \text{Power consumption per GPU in kW}}_{\text{Total power consumed by the model in kW}}$$

- Total number of tokens generated over 1 year:
  nb_user_per_day * nb_request_per_user_per_day * average_token_per_request * total_duration
- Access to column “Output Speed (token/sec)”
- Access to column “Number GPUs Required for Inference”

+ Multiply by carbon intensity

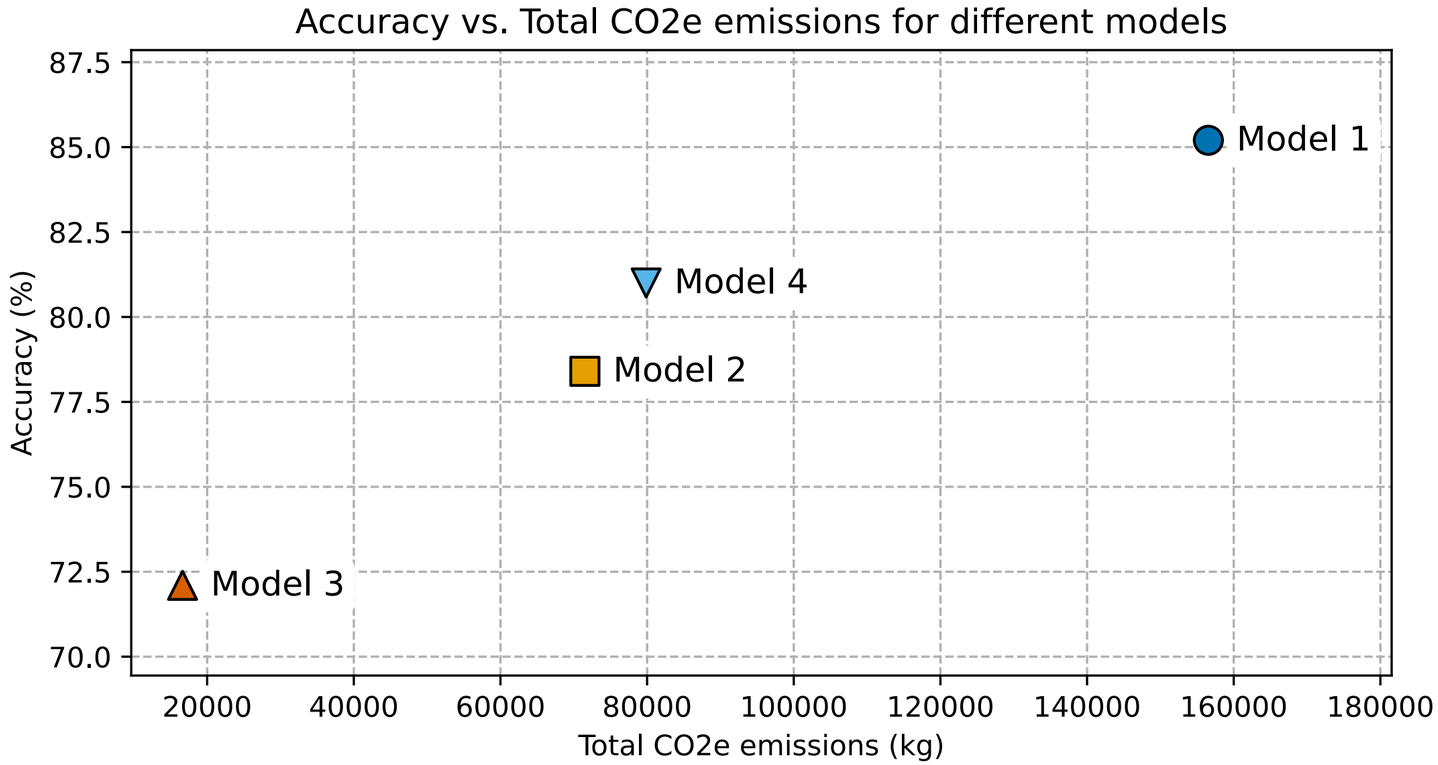
1.2.3 Total emissions

Training CO2e (kg)	Inference CO2e (kg)	Total CO2e (kg)
117000.0	39541.666667	156541.666667
46800.0	24713.541667	71513.541667
11700.0	4942.708333	16642.708333
66690.0	13180.555556	79870.555556

- Total emissions = Training emissions + Inference emissions

1.3.1 Model comparison

Based on the plot above, explain the main strength and weakness for each model: write **1 sentence per model** and **provide numerical evidence** to support your answer. This will be your data to help the ChatCrew company make a decision.



1.3.1 Model comparison

Based on the plot above, explain the main strength and weakness for each model: write **1 sentence per model** and **provide numerical evidence** to support your answer. This will be your data to help the ChatCrew company make a decision.

Example 1: "Model 3 has the lowest total CO2e emissions (~17'000) but also the lowest accuracy (72.1%), so it is the best for the sustainability metric (total CO2e emissions), but the worst for the metric to asses the usefulness of the model (the accuracy). Model 2 has medium CO2e emissions compared to the other models (~70'000) and medium accuracy (78.4%). Model 4 has slightly higher CO2e emissions than model 2 (~80'000) and a better accuracy (81.0%). Finally, Model 4 has the highest C02e emissions (~160'000) but also the highest accuracy (85.2%), so it is the worst based on the sustainability metric, but the best based on the usefulness metric."

Example 2: “model 1: good accuracy but a lot of emissions; model 2: mid accuracy and mid emissions; model 3: low emissions but bad accuracy model; 4: good accuracy and mid emissions”

For each model:

- **Description** of strengths and weaknesses based on accuracy and total CO2e emissions
- **Numbers**

1.3.2 Ethical values & dilemma

What are the **two ethical values / principles** that are opposed in this situation? Which **metrics represent these ethical values / principles** in our analysis?
What is the **dilemma** for the ChatCrew company?
Write 3 sentences.

1.3.2 Ethical values & dilemma – Metrics

Choose 2 metrics used in the notebook that could **correspond to <u>ethical</u> values**:

5% a. Training Time (GPU hours) <- why not but has to be argued

0% b. Model Size (Parameters, in B)

0% c. Output Speed (tokens/sec)

✅ 47% d. Accuracy (%)

0% e. Number GPUs Required for Inference

✅ 47% f. Total CO2e (kg)

0% g. Other

<u>URL:</u> ttpoll.eu
<u>Session ID:</u> cs290

1.3.2 Ethical values & dilemma – Values

Accuracy (%) -> **?**
Total CO2e (kg) -> **?**

URL: ttpoll.eu
Session ID: cs290

Choose 2 ethical values/principles among the ethical values / principles seen in the course:

19% a. Safety
19% b. Fairness
51% c. Sustainability
11% d. Empowerment
0% e. Other

- **Safety:** accuracy reflects errors, which represent negative impacts software can have on its environment
- **Sustainability:** environmental impact reflected by total CO2e
- Empowerment: accuracy reflect errors, which affect end-users (automation bias, hallucinations)
- Fairness: e.g. if accuracy is different for different groups, or considering environmental impacts affecting unfairly different populations, but not really present in the notebook case

Ethical values / principles

Broadly two categories of approaches:

- **Value-oriented methodologies:**
  - Any type of value as defined by stakeholders
  - Human values from established frameworks e.g., Schwartz
- **Principle-based approaches:**
  - Human rights: 30 rights defined by the UDHR
  - Humanitarian principles:
    humanity, neutrality, impartiality and independence
  - Bioethics principles:
    beneficence, non-maleficence, autonomy, and justice
  - **Specific ethical principles for the digital domain?**

Example: Ryan & Stahl, 2020

Principle	Constituent ethical issues or guidance			
Transparency	transparency	explainability	explicability	understandability
	interpretability	communication	disclosure	showing
Justice and fairness	justice	fairness	consistency	inclusion
	equality	equity	non-bias	non-discrimination
	diversity	plurality	accessibility	reversibility
	remedy	redress	challenge	access and distribution
Non-maleficence	non-maleficence	security	safety	harm
	protection	precaution	prevention	integrity
	non-subversion			
Responsibility	responsibility	accountability	liability	acting with integrity
Privacy	privacy	personal or Private information		
Beneficence	benefits	beneficence	well-being	peace
	social good	common good		
Freedom and autonomy	freedom	autonomy	consent	choice
	self-determination	liberty	empowerment	
Trust	trustworthiness			
Sustainability	sustainability	environment (nature)	energy	resources (energy)
Dignity	dignity			
Solidarity	solidarity	social security	cohesion	

Table 1. Guiding ethical principles and constituent ethical issues

1.3.2 Ethical values & dilemma

Example 1: "Here, the two ethical values that are opposed are **sustainability** represented with the CO2e emissions and the **safety** represented with the accuracy. The dilemma for the ChatCrew company is whether it accepts a more accurate model with a worse performance for the environment to prevent false information, or accepts a better model for the environment with a worse performance in terms of accuracy, which then would lead to more false content. Hence we search for the best trade off that does not let performance go off environmental concerns.“

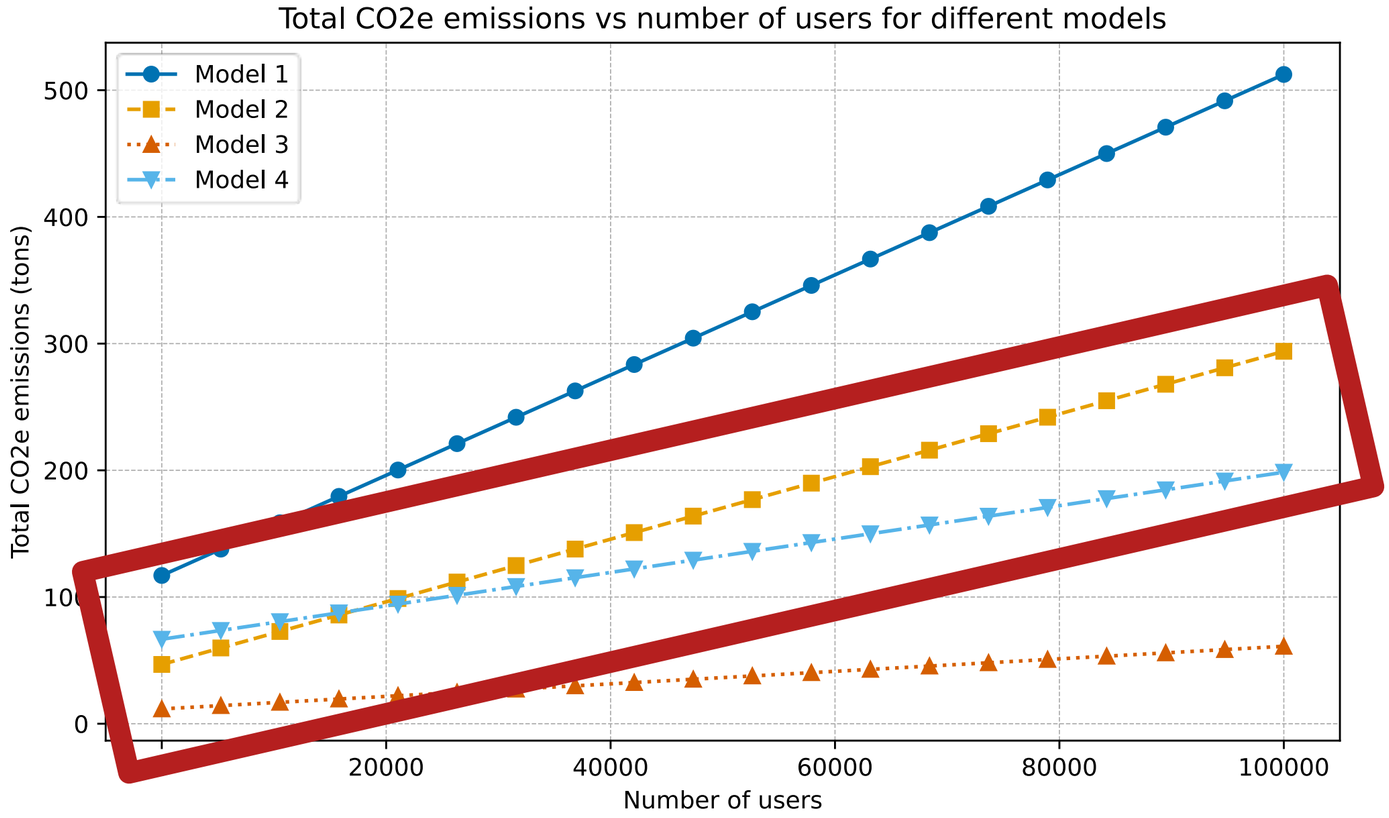
Example 2: "As common in machine learning sustainability cases, we have a focus for sustainability (Universalism-Nature) measured by the carbon footprint opposed with the desire to not mislead or misinform users by giving incorrect results (Benevolence-Dependability and/or Conformity-Rules depending if they state that correct results are their guarantee) measured by accuracy.
The ChatCrew company must thus decide what ethical value they prioritize, and take a decision accordingly. They could for example use a decision matrix using the two previously mentioned metrics as criterion to help them make the decision."

- **2 ethical** values + corresponding **metric**
- **Dilemma**

1.4.1 Influence of number of users

The ChatCrew company has selected **Model 2** and **Model 4** as potential candidates for their product. Given the new expectation of **100 000 users** instead of 10 000 during the year, what should be their final decision?

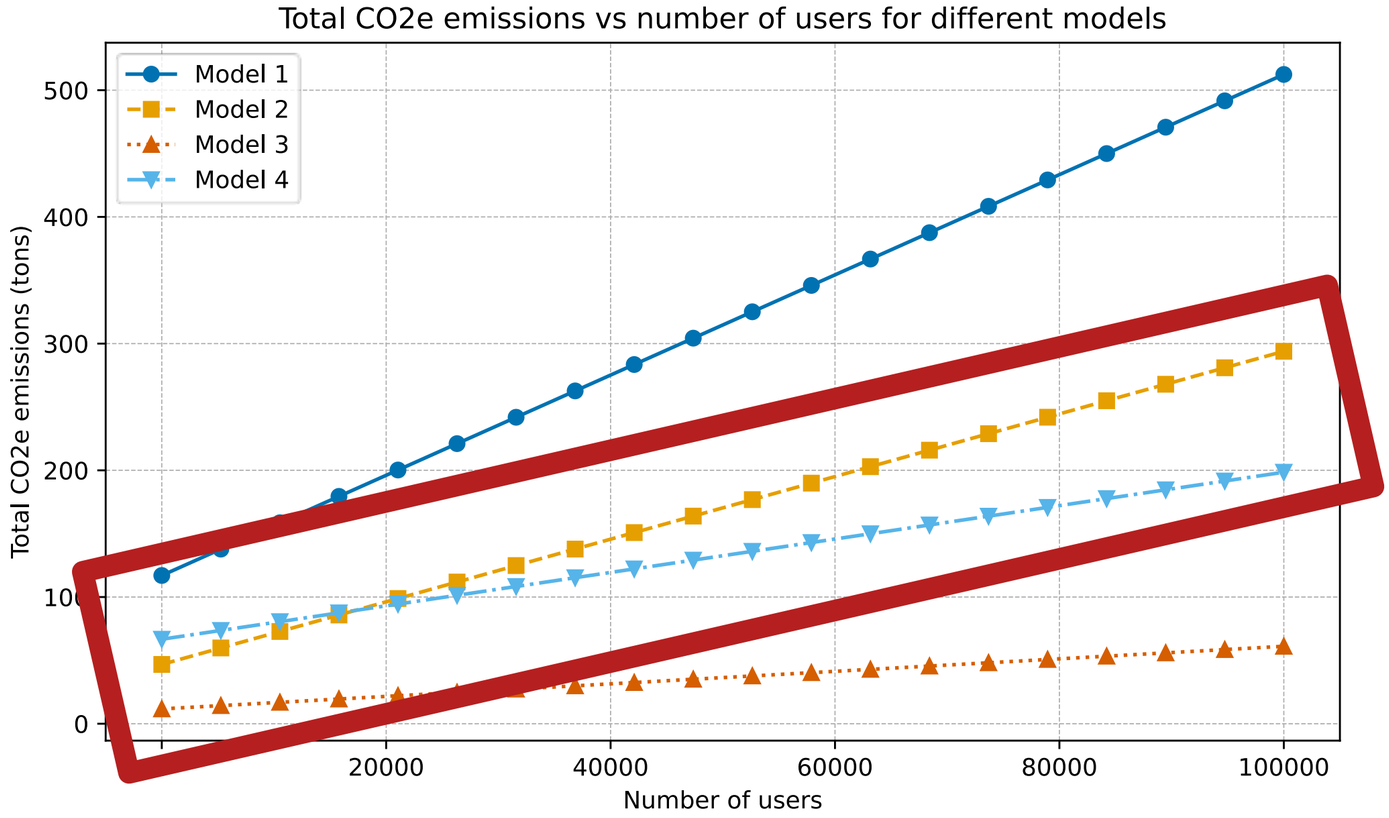
Write 2-3 sentences detailing the **choice** they should make, the **criteria** they should use for that choice, and the corresponding **numerical evidence**.



1.4.1 Influence of number of users

“The final decision should be **model 4**. Because it has a **better accuracy** than model 2 (81.0% vs 78.4%) and the **total CO2 emissions grows slower** as the number of users increase, resulting in a total C02 emission for 100k users that is smaller than the model 2 (200 tons vs 300 tons). So in this case there is no debate, model 4 has a better accuracy for less CO2 emissions than model 2.”

- Final choice: Model 4
- 2 criteria: **accuracy** and **carbon footprint**
- Numerical evidence



Exercise 2

The Carbon Footprint of ChatGPT

Questions which created more difficulty

- **2.1.1 (code) Usage metrics**

- 2.2.1 (code) Carbon footprint per token

- 2.3.1 (code) Variables for EPFL estimation
- 2.3.2 (code) Function for EPFL estimation

- **2.4.1 (open) Embodied emissions**
- 2.4.2 (open) Datacenter upgrade

2.1.1 Usage metrics

	Date	Number of characters	Number of queries
0	2023-09-19	7667	34
1	2023-09-20	3439	4
2	2023-09-21	12296	22
3	2023-09-23	1660	2
4	2023-09-24	7605	8

total_number_queries = epfl_student_df["Number of queries"]**.sum()**
total_character_year = epfl_student_df["Number of characters"]**.sum()**
total_token_year = total_character_year / 4
average_query_day = total_number_queries **/ 365**
average_query_size = **total_character_year / total_number_queries**

2.2.1 Carbon footprint per token

$$\text{Electricity for 1 token generated} = \underbrace{\frac{1\ \text{Token}}{\text{Number of token generated per second}}}_{\text{Time in seconds to generate one token}} \times \underbrace{\frac{1}{3600}}_{\text{Conversion in hours}} \times \underbrace{\text{Number of GPUs} \times \text{Power consumption per GPU in kW} \times \text{PUE}}_{\text{Total power consumed by the model in kW}}$$

time_per_token_seconds = 1 / 107.5
time_per_token_hours = time_per_token_seconds / 3600
power_consumed = 8 * **0.407** * 1.2
footprint_per_token = time_per_token_hours * power_consumed * **262**

👉The amount of CO2e emitted per token is 0.0026 grams of CO2e.

- Power consumption per GPU in **kW**
- Carbon intensity in **g CO2e / kWh**

2.3.1 Variables for EPFL estimation

number_students = 10000 * 0.75

token_per_student_per_year = total_token_year

lower_bound_carbon_emissions_kg_per_token = 0.0008 **/ 1000**
upper_bound_carbon_emissions_kg_per_token = footprint_per_token **/ 1000**

- Emissions per token in **kg CO2e / kWh**

2.3.2 Function for EPFL estimation

Let's complete the following function yearly_carbon_emissions_chatgpt() to calculate the yearly estimate of the carbon footprint from ChatGPT for a given number of people, a given number of tokens generated per year and a given emissions per token. The function should **return** the yearly carbon footprint in **kg CO2e**.
The parameters of the function are the following:
- total_number_people : the total number of people using ChatGPT
- total_number_tokens : the total number of tokens generated by a single person during the year
- emissions_per_token : the emissions per token generated by ChatGPT in **kg CO2e**

carbon_emissions_per_year =
total_number_people * total_number_tokens * emissions_per_token

- No conversion needed in the function

EPFL carbon footprint estimation

For the lower bound, the yearly carbon emissions of ChatGPT are:

7 016 kg of CO_2eq are equivalent to:
🚙 driving an average passenger car for 42 807 km,
🛫 taking an international flight for 37 737 km or
🚆 travelling by train for 1 573 100 km.

Lower bound: 0,0008 g CO2e / token

For the upper bound, the yearly carbon emissions of ChatGPT are:

23 198 kg of CO_2eq are equivalent to:
🚙 driving an average passenger car for 141 540 km,
🛫 taking an international flight for 124 776 km or
🚆 travelling by train for 5 201 426 km.

Upper bound: 0,0026 g CO2e / token

In **carbon budget**, where the target is 2 t CO2e per person per year:

- Lower bound = 3,5 person-years per year
- Upper bound = 11,6 person-years per year

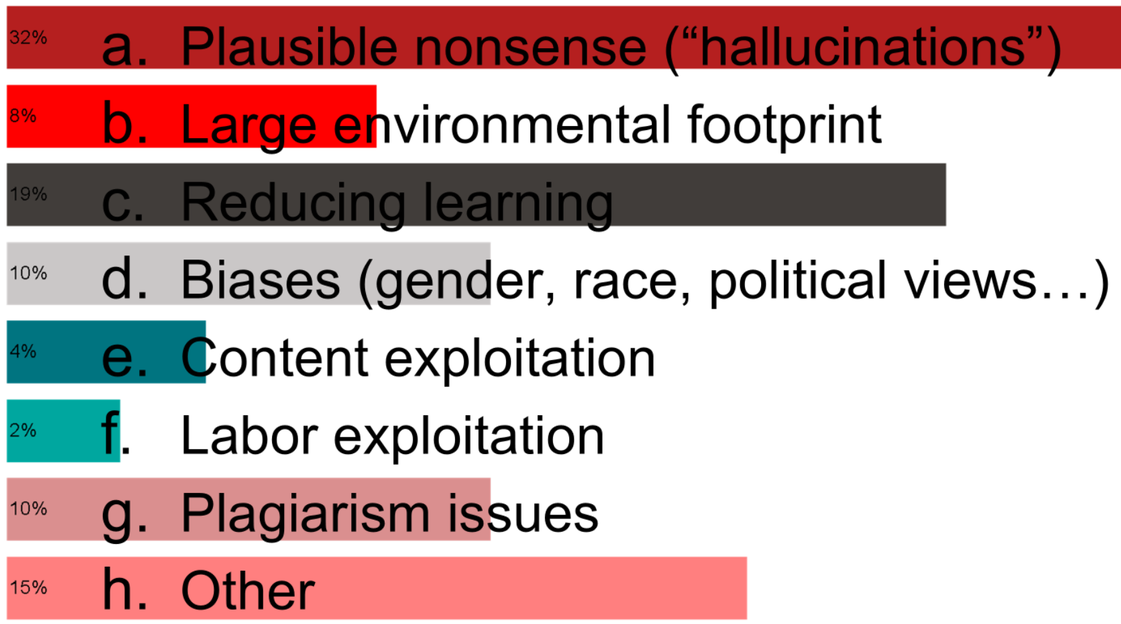
Most recent estimation (Verma & Tan, 2024):

- 0,262 g CO2e / token
- **1 148 person-years** per year // 15% of each student's yearly budget

EPFL carbon footprint estimation

What are the drawbacks (risks) of Generative AI?

Select all that apply:



2.4.1 Embodied emissions

What do the embodied emissions of a model represent in the case of ChatGPT? Write 3 sentences to explain this concept.

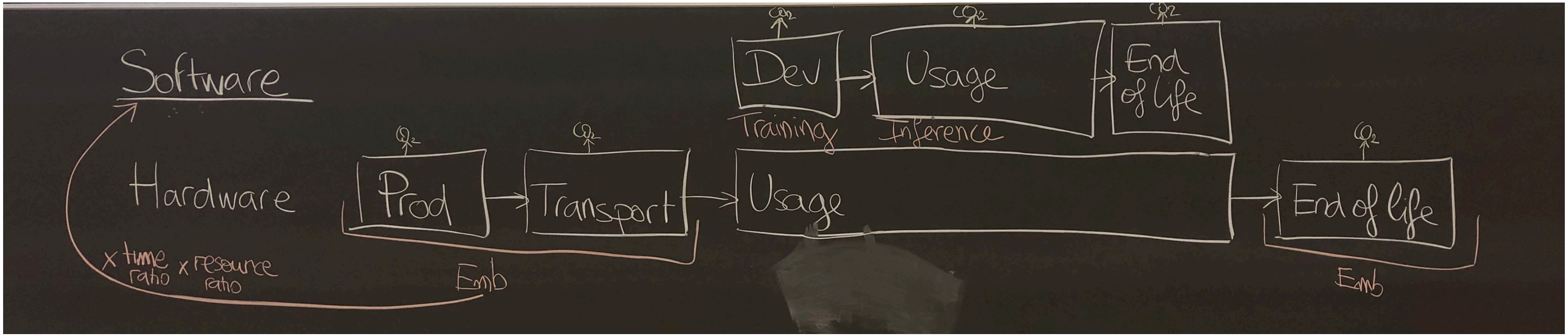
"As seen in the course, embodied emissions correspond to emissions associated with production/transport/end of life phases of a product, so usage is not considered here (inference). For software, we will calculate embodied emissions by multiplying the embodied emissions of the considered hardware by a time share factor(Execution time/hardware lifespan) and by a ressource share factor(Execution ressources / Hardware ressources). For chat gpt software, hardware would be servers and GPUs, and time factors things like training time."

- **Hardware** components, mostly integrated circuits such as GPUs, CPUs, RAM…
- Phases **other than use**, mostly manufacturing but also transport and EOL
- A **share** is attributed to software depending on time and resources used
  (for ML models, both training & inference time & resources should be “counted”)

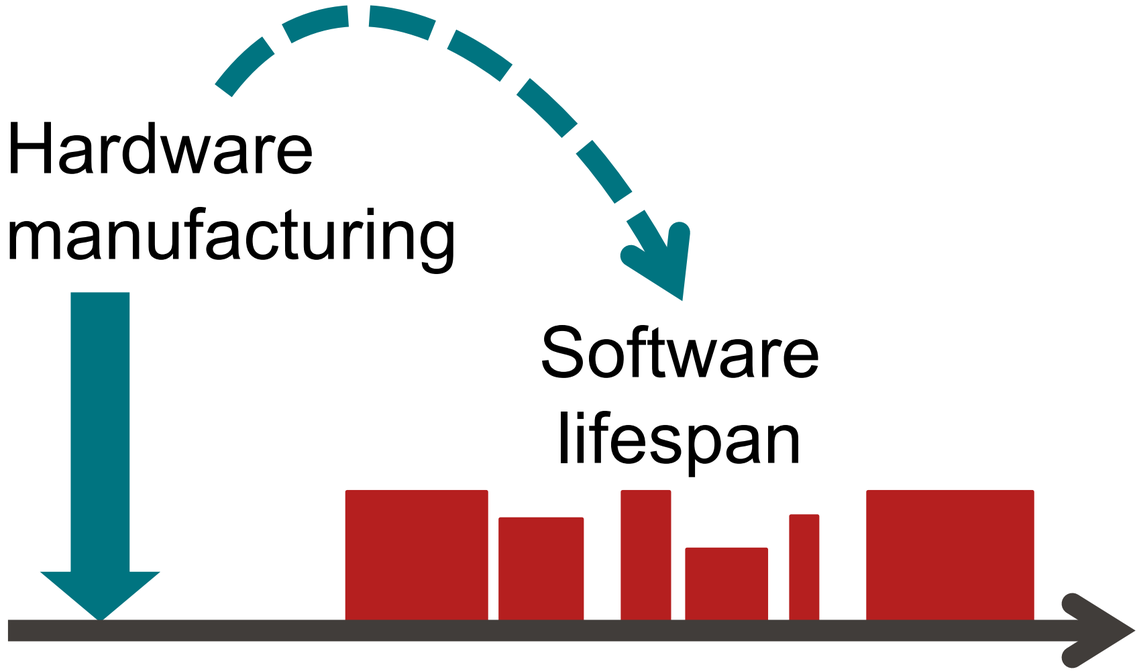
Clarification on Embodied Emissions

- Software is not a physical object, so per se it does not have “embodied emissions” since it does not have a body
- BUT software does not work without hardware, and hardware has embodied emissions

👉 A share of the embodied emissions of hardware is attributed to software, depending on how much the hardware is used by software



Attributing embodied emissions to software



$$\boldsymbol{M_S} = M_H \times R_T \times R_R$$

- M_H: Embodied emissions of hardware
- R_T: **Time share** = Execution time / Hardware lifespan
- R_R: **Resource share** = Execution resources / Hardware resources

👉**Share** based on:

- Execution time
- Used resource

The example of BLOOM

(Morand et al., 2024; Luccioni et al., 2024)

- Hardware:
  - 48 servers,
    2500 kg CO2e / server
  - 385 GPUs,
    150 kg CO2e / GPU
  - Lifespan: 44 676 hours
    (6 years used at 85%)
- Software:
  - Training time: 2 820 hours
  - Resources used: 100%

👉 **Attributed embodied emissions:**

$$\boldsymbol{M_S} = (120\,000 + 57\,750) \times \frac{2\,820}{44\,676}$$

$$\boldsymbol{M_S = 11.2}\ \text{t}\ \text{CO}_2\text{e}$$

In addition to emissions from training itself:
24.7 t CO2e

2.4.2 Datacenter upgrade

Let's imagine that, after two years of use, OpenAI wants to upgrade the GPUs used in their datacenter(s). In an effort to improve ChatGPT's sustainability, they choose GPUs that are twice as fast as the previous ones in terms of computing speed, for the exact same energy consumption.

“This would reduce the amount of time to generate a token and thus lead to lower power consumption for the same task. However, the embodied cost of the GPUs needs to be considered to. If the renewal of the GPUs is too frequent, **the savings in emissions from more efficient GPU is offset by the embodied emissions of the GPUs**.
It is also possible for this change to **induce demand**. If the computation is noticeably faster from the users perspective they might use more requests. This could also offset the energy consumption because of a higher volume of queries. Finally this is under the assumption that the GPU's uptime can be allocated optimally to take advantage of the lower computation time. While idle both architectures have the same power consumption.”

- **Embodied emissions**
- **Rebound effect**

But also:
- E-waste
- Lifespan too short
- Idle time
- Cooling
- …

Conclusion

When computing sustainability metrics:

- Computation are not very complex but…
- Units are always a pain
- Numbers often difficult to make concrete

What’s next?

Next dates

	Monday (SG1)	Tuesday (Computer Rooms)
2 Dec – 8 Dec	Debriefing Graded 2	Empowerment 2 notebook
9 Dec – 15 Dec	Empowerment 2 cases	Conclusion & Q&A in **SG1**
16 Dec – 20 Dec	Final exam	---

- Empowerment 2:
  - 1 notebook
  - Only **1 video** + quizzes
  - Review cases (bad actors, ethical speculation, datasheet) + review quiz
- Conclusion:
  - Review cases (digital ethics canvas, ethics canvas)
  - Q & A

Prep for Q&A session

Before December 9 at 10h, post on SpeakUp the things you would like to discuss on December 10, 8h15-10h:

- Course content
- Quizzes
- Case studies
- Strategies

👉 1 post = 1 question/ideas

👍 Vote for other's ideas

Post your ideas:
https://speakup.epfl.ch
Room key: **53228**

References

- Ryan, M., & Stahl, B. C. (2020). Artificial intelligence ethics guidelines for developers and users: Clarifying their content and normative implications. Journal of Information, Communication and Ethics in Society, 19(1), 61–86. https://doi.org/10.1108/JICES-12-2019-0138

- A bottle of water per email: The hidden environmental costs of using AI chatbots. (2024, September 18). Washington Post. https://www.washingtonpost.com/technology/2024/09/18/energy-ai-use-electricity-water-data-centers/