

Natural Language Processing

# Developing LLMs

Marco Kuhlmann

Department of Computer and Information Science

# This session

- Survey & announcements
- Developing LLMs (Q&A)
- Ethical issues (in-class assignment)
- Introduction to Lab 3

# Survey: The first two weeks

14 responses submitted

What is your overall impression of the course at this point? (1: very bad – 5: very good)

4.50



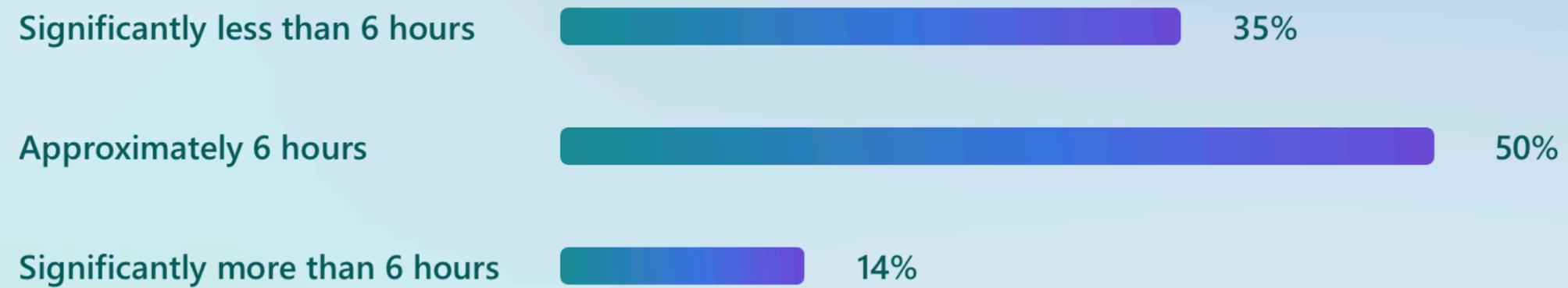
14 responses submitted

## How much time did you spend watching and reflecting on the video lectures for Unit 1 (tokenisation and embeddings)?



14 responses submitted

## How much time did you spend watching and reflecting on the video lectures for Unit 2 (LLM architectures)?



14 responses submitted

## How much time did you spend on lab 1 (tokenisation and embeddings)?



How much time did you spend on each lab?

Is that a good thing or a bad thing?

What feedback did you ask for? Did it help?

How did you use AI tools? Did it help?

# Announcements

- General review of the portfolio
- No project seminar this week
- Pitch your project ideas to me!

# Overview of Unit 3: Developing LLMs

# Overview of this unit

- 3.1 Introduction to LLM development
- 3.2 Training LLMs
- 3.3 Data for LLM pretraining
- 3.4 Scaling laws
- 3.5 Emergent abilities of LLMs
- 3.6 Environmental cost of LLMs

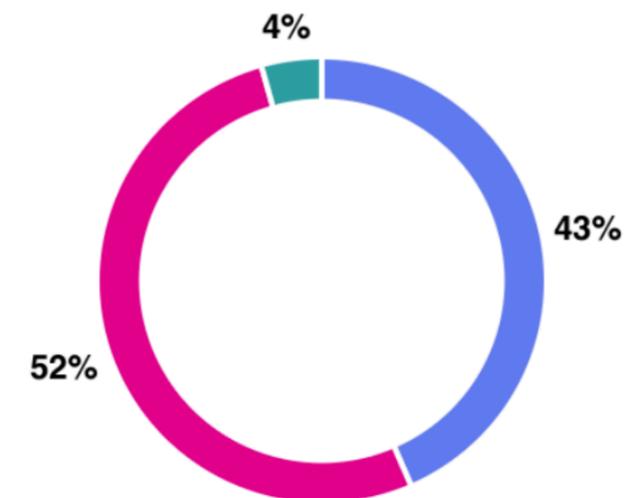
# Lecture 3.2, question 2

2. What is the total norm as referred to in the context of gradient clipping? (1 point)

[More details](#)

43% of respondents answered this question correctly.

- The norm of the vector containing all parameter norms. 10 ✓
- The norm of the vector containing all parameters. 12
- The sum of the norms of all parameter vectors. 1



# torch.nn.utils.clip\_grad\_norm\_

```
torch.nn.utils.clip_grad_norm_(parameters, max_norm, norm_type=2.0,  
error_if_nonfinite=False, foreach=None)
```

[\[source\]](#)

Clip the gradient norm of an iterable of parameters.

The norm is computed over the norms of the individual gradients of all parameters, as if the norms of the individual gradients were concatenated into a single vector. Gradients are modified in-place.

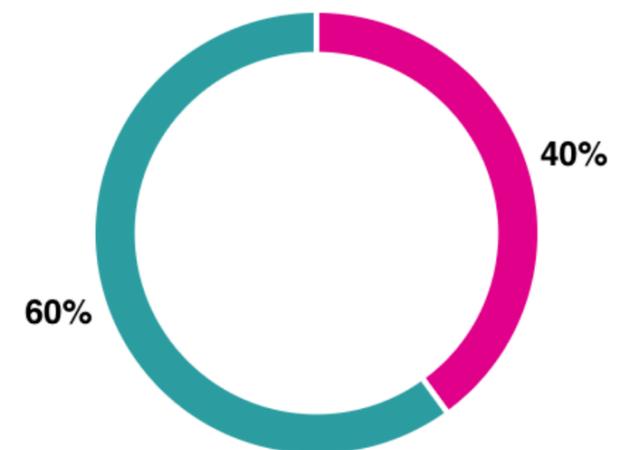
# Lecture 3.2, question 3

3. In the learning rate scheduler example, what is the learning rate after 280B tokens? (1 point)

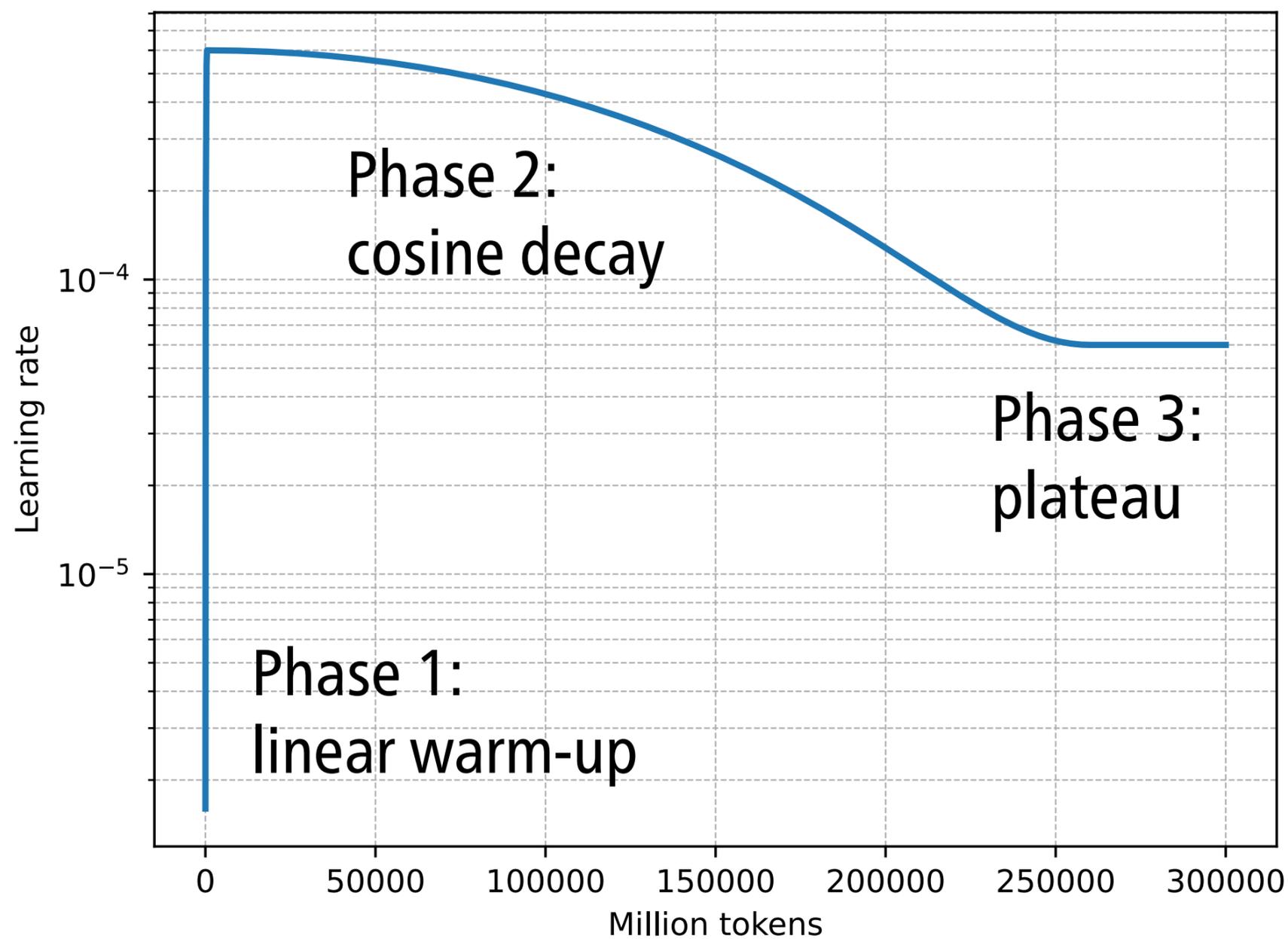
[More details](#)

60% of respondents answered this question correctly.

● 0.006	0
● 0.0006	10
● 0.00006	15 ✓



# Learning rate scheduling



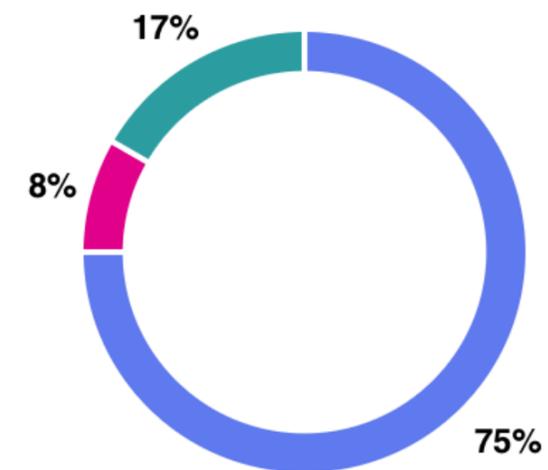
# Lecture 3.3, question 5

5. For the largest ablation model considered, how many points of average accuracy is FineWeb-EDU better than standard FineWeb? (1 point)

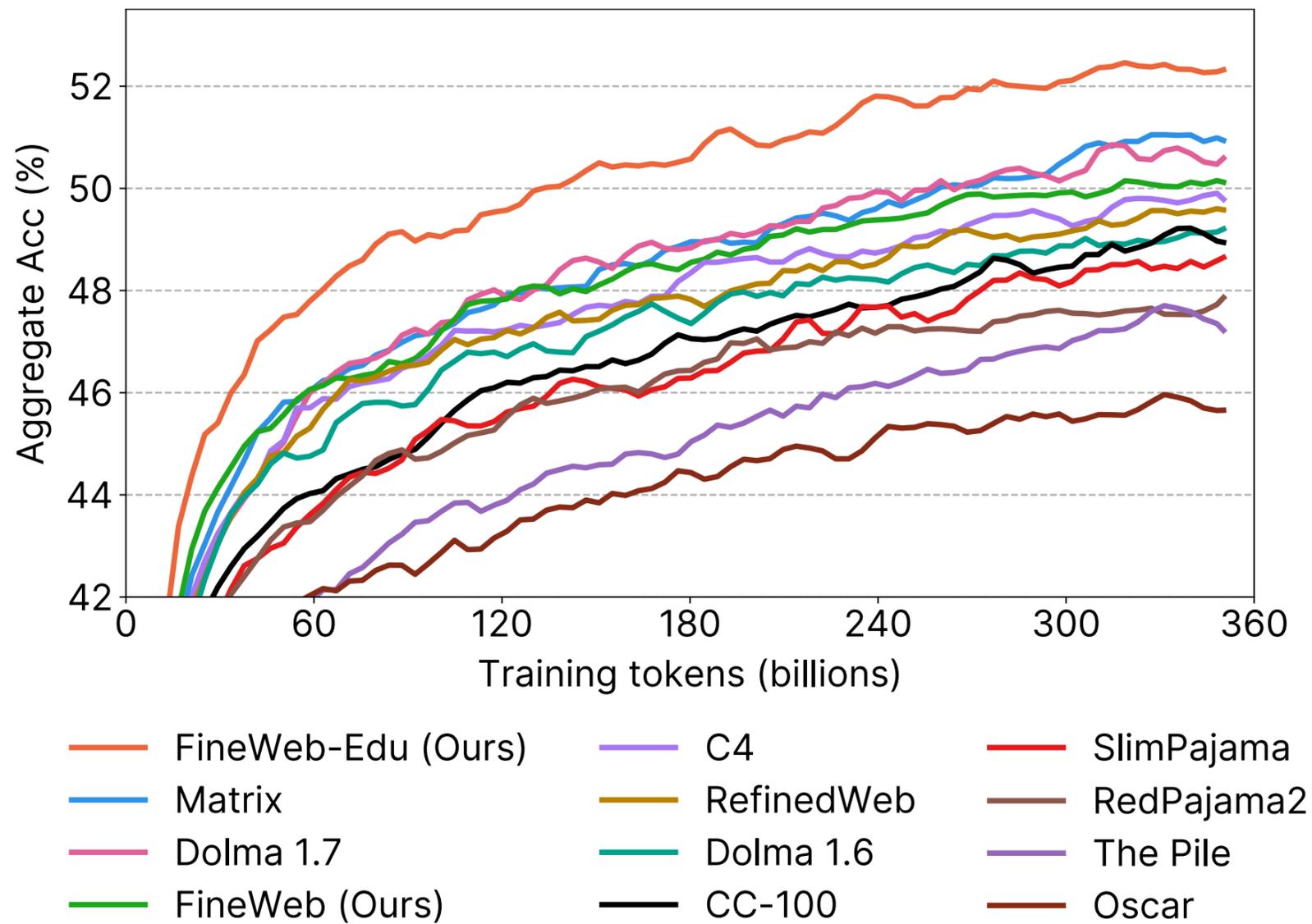
[More details](#)

75% of respondents answered this question correctly.

● 2 points	18 ✓
● 4 points	2
● 6 points	4



# Impact of high-quality data



Penedo et al. (2024)

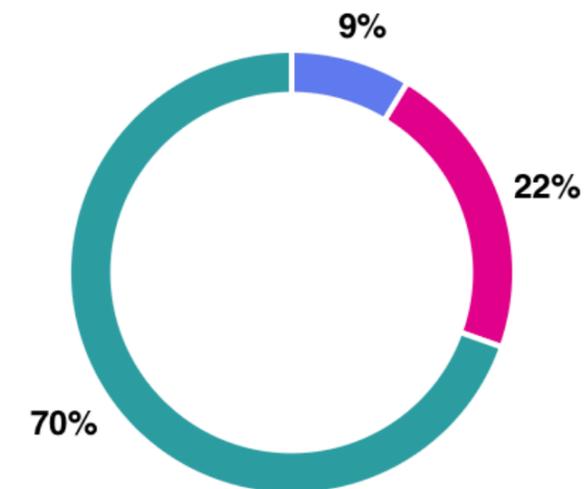
# Lecture 3.4, question 5

5. Based on the results from the Chinchilla paper, which of the following models was least compute-optimal? (1 point)

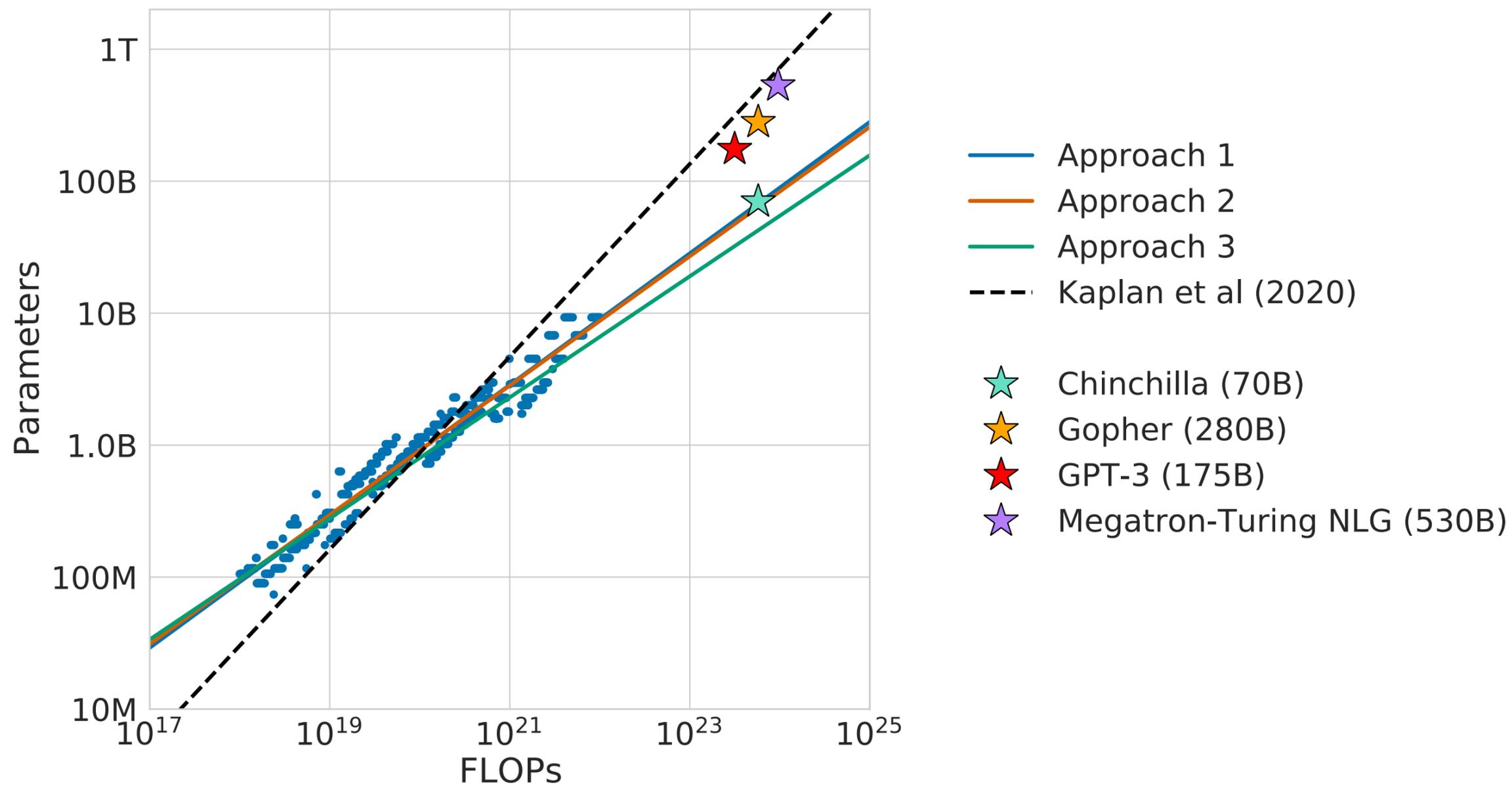
[More details](#)

70% of respondents answered this question correctly.

● Gopher (280B)	2
● GPT-3 (175B)	5
● Megatron-Turing NLG (530B)	16 ✓



# Large language models can be too large



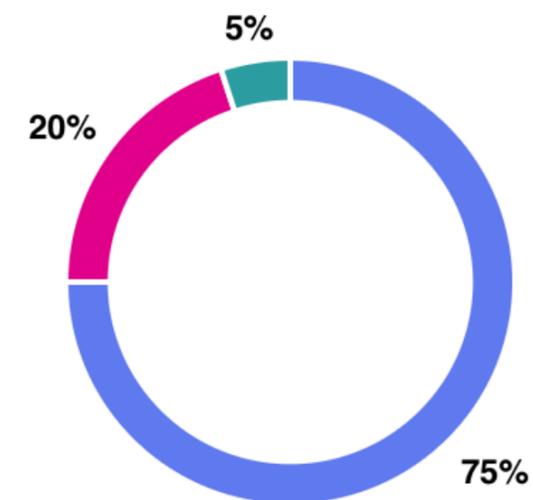
# Lecture 3.5, question 5

5. In the slide on prompt engineering, how much better was the LLM-designed prompt compared to the best human-designed prompt? (1 point)

[More details](#)

75% of respondents answered this question correctly.

● 3.3 points	15 ✓
● 3.3 percent	4
● 3.3 times	1



# Prompt engineering

designed by a LM



Prompt	Accuracy
Let's work this out in a step by step way to be sure we have the right answer.	<b>82.0</b>
Let's think step by step.	78.7
First, ...	77.3
Let's think about this logically.	74.5
Let's solve this problem by splitting it into steps	72.2
Let's be realistic and think step by step.	70.8
Let's think like a detective step by step.	70.3
(Zero-shot)	17.7

**Ethical issues**

BUSINESS • TECHNOLOGY

# Exclusive: OpenAI Used Kenyan Workers on Less Than \$2 Per Hour to Make ChatGPT Less Toxic



This image was generated by OpenAI's image-generation software, Dall-E 2. The prompt was: "A seemingly endless view of African workers at desks in front of computer screens in a printmaking style." TIME does not typically use AI-generated art to illustrate its stories, but chose to in this instance in order to draw attention to the power of OpenAI's technology and shed light on the labor that makes it possible. Image generated by Dall-E 2/OpenAI

BY **BILLY PERRIGO**  JANUARY 18, 2023 7:00 AM EST



# Tasks

- Read the article.
- Use an AI assistant to generate a 200-word summary.  
*Enter the summary into the form.*
- Discuss with your neighbour:
  - Was the company's behaviour ethical?
  - Do you consider the AI-generated summary a good summary?*Enter your answer to the latter question into the form.*



<https://forms.office.com/e/FHaLGJ1Nt2>

# Introduction to Lab L3