Large Language Model Fundamentals

Bo Kang

Ghent University

Outline

- Introduction to Large Language Models
- Training an LLM: Behind the Scenes
- Understanding the Costs
- Hands-on: Deploy Your Own Coding Copilot

Introduction to Large Language Models

- An AI model that predicts the next word or phrase in text.
- Learns patterns by processing massive amounts of written text.
- Requires significant computational resources to train, store, and run.
- Capable of complex language tasks (e.g., answering questions, summarizing, coding assistance).

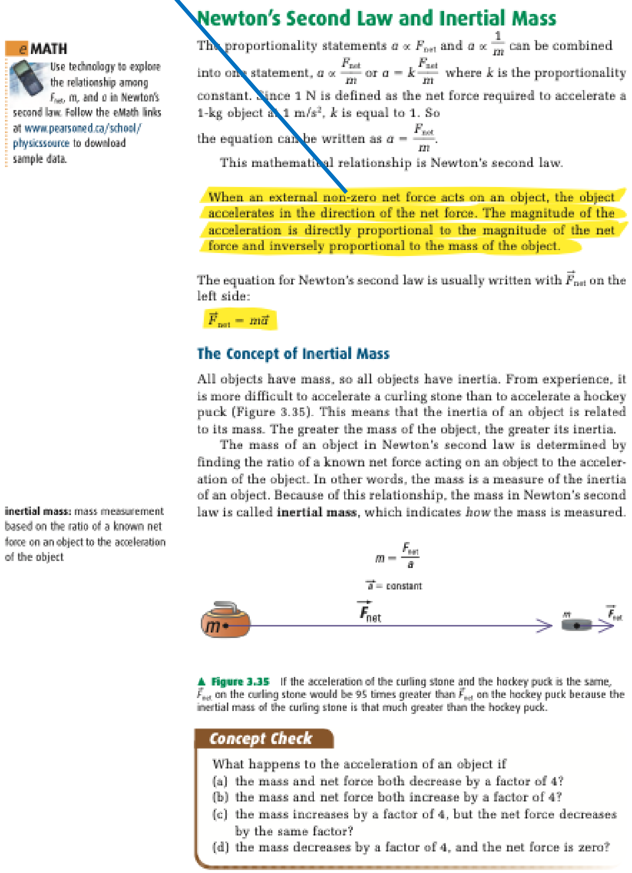
Outline

- Introduction to Large Language Models
- Training an LLM: Behind the Scenes
- Understanding the Costs
- Hands-on: Deploy Your Own Coding Copilot

Overview: analogy of studying a textbook

Exposition ⇔ Pretraining
(background knowledge)

Worked problems ⇔ Supervised Finetuning
(problem + demonstrated solution, for imitation)


eMATH
Use technology to explore the relationship among F_{net}, m, and a in Newton's second law. Follow the eMath links at www.pearsoned.ca/school/physicssource to download sample data.

Newton's Second Law and Inertial Mass

The proportionality statements $a \propto F_{net}$ and $a \propto \frac{1}{m}$ can be combined into one statement, $a \propto \frac{F_{net}}{m}$ or $a = k\frac{F_{net}}{m}$ where k is the proportionality constant. Since 1 N is defined as the net force required to accelerate a 1-kg object at 1 m/s², k is equal to 1. So the equation can be written as $a = \frac{F_{net}}{m}$.

This mathematical relationship is Newton's second law.

When an external non-zero net force acts on an object, the object accelerates in the direction of the net force. The magnitude of the acceleration is directly proportional to the magnitude of the net force and inversely proportional to the mass of the object.

The equation for Newton's second law is usually written with $\vec{F}_{net}$ on the left side:

$\vec{F}_{net} = m\vec{a}$

The Concept of Inertial Mass

All objects have mass, so all objects have inertia. From experience, it is more difficult to accelerate a curling stone than to accelerate a hockey puck (Figure 3.35). This means that the inertia of an object is related to its mass. The greater the mass of the object, the greater its inertia.

The mass of an object in Newton's second law is determined by finding the ratio of a known net force acting on an object to the acceleration of the object. In other words, the mass is a measure of the inertia of an object. Because of this relationship, the mass in Newton's second law is called **inertial mass**, which indicates *how* the mass is measured.

inertial mass: mass measurement based on the ratio of a known net force on an object to the acceleration of the object

$m = \frac{F_{net}}{a}$

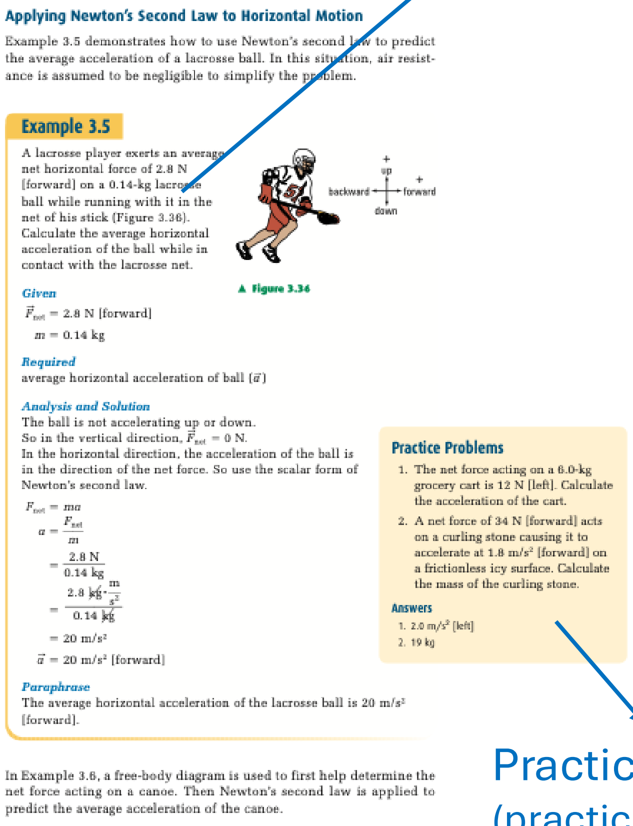
$\vec{a}$ = constant

$\vec{F}_{net}$

▲ **Figure 3.35** If the acceleration of the curling stone and the hockey puck is the same, $\vec{F}_{net}$ on the curling stone would be 95 times greater than $\vec{F}_{net}$ on the hockey puck because the inertial mass of the curling stone is that much greater than the hockey puck.

Concept Check

What happens to the acceleration of an object if
(a) the mass and net force both decrease by a factor of 4?
(b) the mass and net force both increase by a factor of 4?
(c) the mass increases by a factor of 4, but the net force decreases by the same factor?
(d) the mass decreases by a factor of 4, and the net force is zero?


Applying Newton's Second Law to Horizontal Motion

Example 3.5 demonstrates how to use Newton's second law to predict the average acceleration of a lacrosse ball. In this situation, air resistance is assumed to be negligible to simplify the problem.

Example 3.5

A lacrosse player exerts an average net horizontal force of 2.8 N [forward] on a 0.14-kg lacrosse ball while running with it in the net of his stick (Figure 3.36). Calculate the average horizontal acceleration of the ball while in contact with the lacrosse net.

backward forward up down

▲ Figure 3.36

Given
$\vec{F}_{net}$ = 2.8 N [forward]
m = 0.14 kg

Required
average horizontal acceleration of ball ($\vec{a}$)

Analysis and Solution
The ball is not accelerating up or down.
So in the vertical direction, $\vec{F}_{net}$ = 0 N.
In the horizontal direction, the acceleration of the ball is in the direction of the net force. So use the scalar form of Newton's second law.

$$F_{net} = ma$$
$$a = \frac{F_{net}}{m} = \frac{2.8\ \text{N}}{0.14\ \text{kg}} = \frac{2.8\ \cancel{\text{kg}}\cdot\frac{\text{m}}{\text{s}^2}}{0.14\ \cancel{\text{kg}}} = 20\ \text{m/s}^2$$
$$\vec{a} = 20\ \text{m/s}^2\ [\text{forward}]$$

Paraphrase
The average horizontal acceleration of the lacrosse ball is 20 m/s² [forward].

Practice Problems
1. The net force acting on a 6.0-kg grocery cart is 12 N [left]. Calculate the acceleration of the cart.
2. A net force of 34 N [forward] acts on a curling stone causing it to accelerate at 1.8 m/s² [forward] on a frictionless icy surface. Calculate the mass of the curling stone.

Answers
1. 2.0 m/s² [left]
2. 19 kg

In Example 3.6, a free-body diagram is used to first help determine the net force acting on a canoe. Then Newton's second law is applied to predict the average acceleration of the canoe.

Practice Problems ⇔ Reinforcement learning
(practice, trial & error until reach correct answer)

Overview: analogy of studying a textbook

Exposition ⇔ Pretraining
(background knowledge)

Worked problems ⇔ Supervised Finetuning
(problem + demonstrated solution, for imitation)

eMATH

Use technology to explore the relationship among F_{net}, m, and a in Newton's second law. Follow the eMath links at www.pearsoned.ca/school/physicssource to download sample data.

Newton's Second Law and Inertial Mass

The proportionality statements $a \propto F_{net}$ and $a \propto \frac{1}{m}$ can be combined into one statement, $a \propto \frac{F_{net}}{m}$ or $a = k\frac{F_{net}}{m}$ where k is the proportionality constant. Since 1 N is defined as the net force required to accelerate a 1-kg object at 1 m/s^2, k is equal to 1. So the equation can be written as $a = \frac{F_{net}}{m}$.

This mathematical relationship is Newton's second law.

When an external non-zero net force acts on an object, the object accelerates in the direction of the net force. The magnitude of the acceleration is directly proportional to the magnitude of the net force and inversely proportional to the mass of the object.

The equation for Newton's second law is usually written with $\vec{F}_{net}$ on the left side:

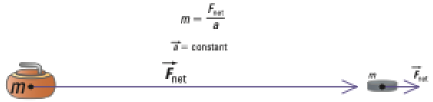
$$\vec{F}_{net} = m\vec{a}$$

The Concept of Inertial Mass

All objects have mass, so all objects have inertia. From experience, it is more difficult to accelerate a curling stone than to accelerate a hockey puck (Figure 3.35). This means that the inertia of an object is related to its mass. The greater the mass of the object, the greater its inertia.

The mass of an object in Newton's second law is determined by finding the ratio of a known net force acting on an object to the acceleration of the object. In other words, the mass is a measure of the inertia of an object. Because of this relationship, the mass in Newton's second law is called **inertial mass**, which indicates *how* the mass is measured.

inertial mass: mass measurement based on the ratio of a known net force on an object to the acceleration of the object



▲ **Figure 3.35** If the acceleration of the curling stone and the hockey puck is the same, $\vec{F}_{net}$ on the curling stone would be 95 times greater than $\vec{F}_{net}$ on the hockey puck because the inertial mass of the curling stone is that much greater than the hockey puck.

Concept Check

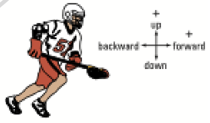
What happens to the acceleration of an object if
(a) the mass and net force both decrease by a factor of 4?
(b) the mass and net force both increase by a factor of 4?
(c) the mass increases by a factor of 4, but the net force decreases by the same factor?
(d) the mass decreases by a factor of 4, and the net force is zero?

Applying Newton's Second Law to Horizontal Motion

Example 3.5 demonstrates how to use Newton's second law to predict the average acceleration of a lacrosse ball. In this situation, air resistance is assumed to be negligible to simplify the problem.

Example 3.5

A lacrosse player exerts an average net horizontal force of 2.8 N [forward] on a 0.14-kg lacrosse ball while running with it in the net of his stick (Figure 3.36). Calculate the average horizontal acceleration of the ball while in contact with the lacrosse net.



▲ **Figure 3.36**

Given
$\vec{F}_{net}$ = 2.8 N [forward]
m = 0.14 kg

Required
average horizontal acceleration of ball ($\vec{a}$)

Analysis and Solution
The ball is not accelerating up or down.
So in the vertical direction, $\vec{F}_{net}$ = 0 N.
In the horizontal direction, the acceleration of the ball is in the direction of the net force. So use the scalar form of Newton's second law.

$$\begin{aligned} F_{net} &= ma \\ a &= \frac{F_{net}}{m} \\ &= \frac{2.8\ \text{N}}{0.14\ \text{kg}} \\ &= \frac{2.8\ \cancel{\text{kg}} \cdot \frac{\text{m}}{\text{s}^2}}{0.14\ \cancel{\text{kg}}} \\ &= 20\ \text{m/s}^2 \\ \vec{a} &= 20\ \text{m/s}^2\ [\text{forward}] \end{aligned}$$

Paraphrase
The average horizontal acceleration of the lacrosse ball is 20 m/s^2 [forward].

Practice Problems

1. The net force acting on a 6.0-kg grocery cart is 12 N [left]. Calculate the acceleration of the cart.
2. A net force of 34 N [forward] acts on a curling stone causing it to accelerate at 1.8 m/s^2 [forward] on a frictionless icy surface. Calculate the mass of the curling stone.

Answers

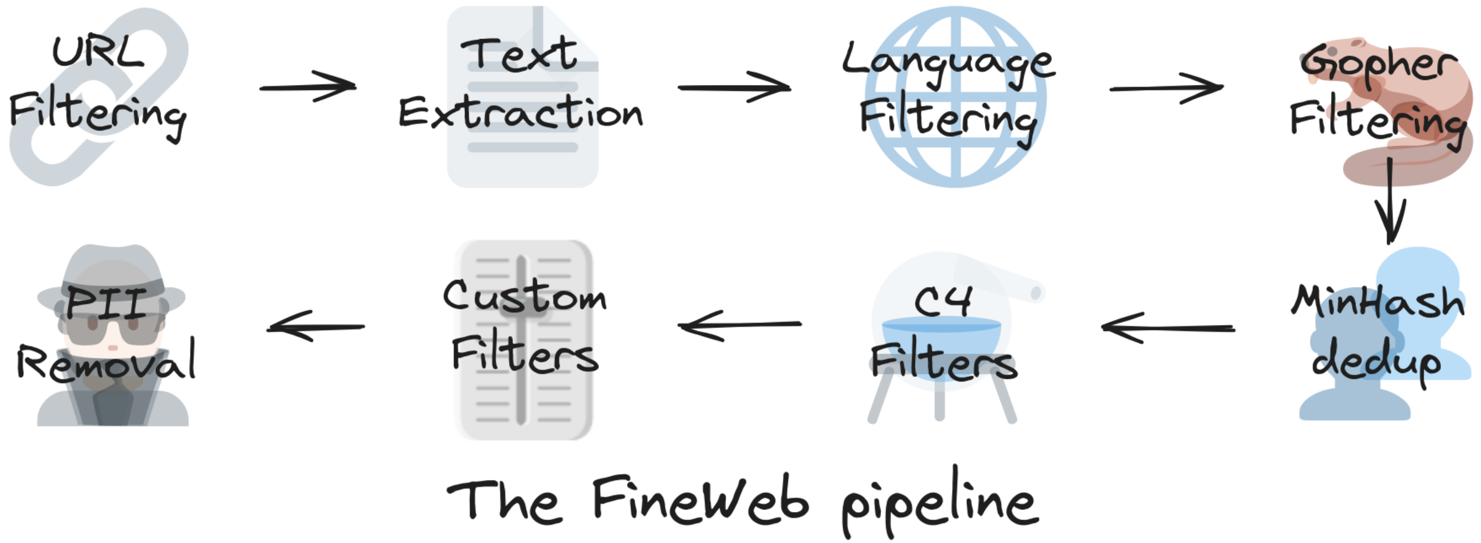
1. 2.0 m/s^2 [left]
2. 19 kg

In Example 3.6, a free-body diagram is used to first help determine the net force acting on a canoe. Then Newton's second law is applied to predict the average acceleration of the canoe.

Practice Problems ⇔ Reinforcement learning
(practice, trial & error until reach correct answer)

Data Collection: Turning the internet into training data

- Recipe



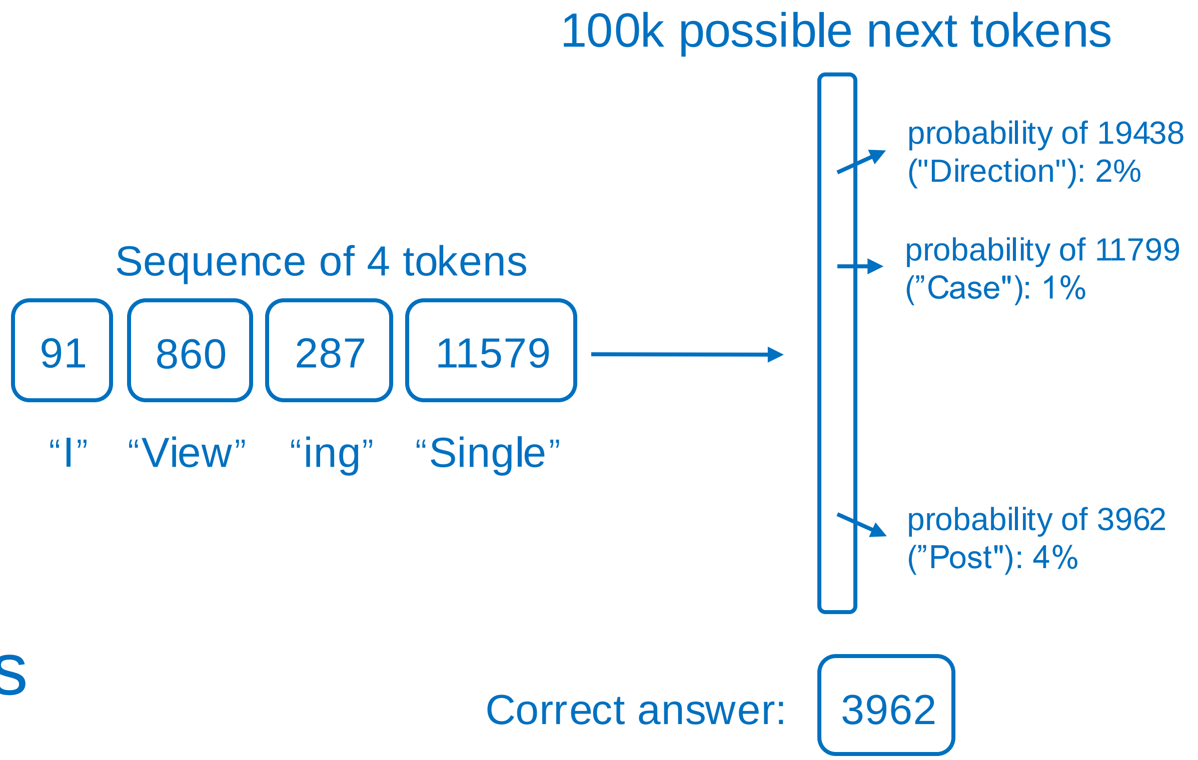
- Examples

Tokenization: making text readable to LLMs

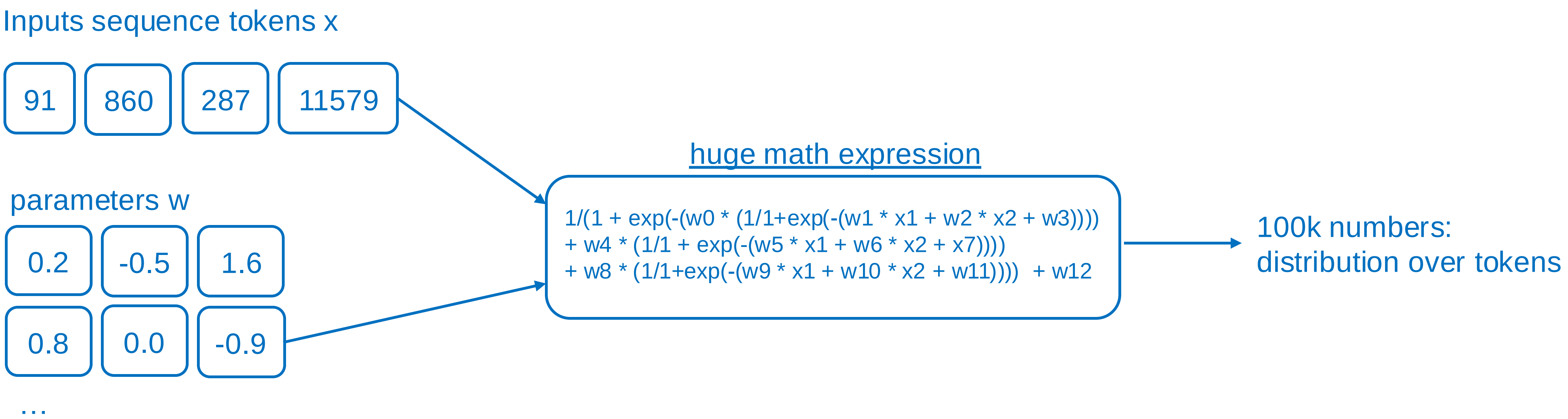
- Computers and LLMs can only "see" and process symbols from a finite set.
- Need to convert text into symbols the LLM understands
- Example: Representing ~5000 text characters
  - ~40,000 bits (vocabulary size of 2, inefficient patterns)
  - ~5000 bytes (vocabulary size of 256, limited patters)
  - ~1300 tokens (vocabulary size of 100k, meaningful patterns, easy for LLM to learn)

Neural network training: learning patterns in tokens

- Goal: Teach the neural network (NN) to predict the next token based on previous ones.
- How it works:
  - Take a sequence of tokens
  - Predict what token likely comes next
  - Compare prediction to the actual next token
  - Adjust NN parameters, improving future predictions



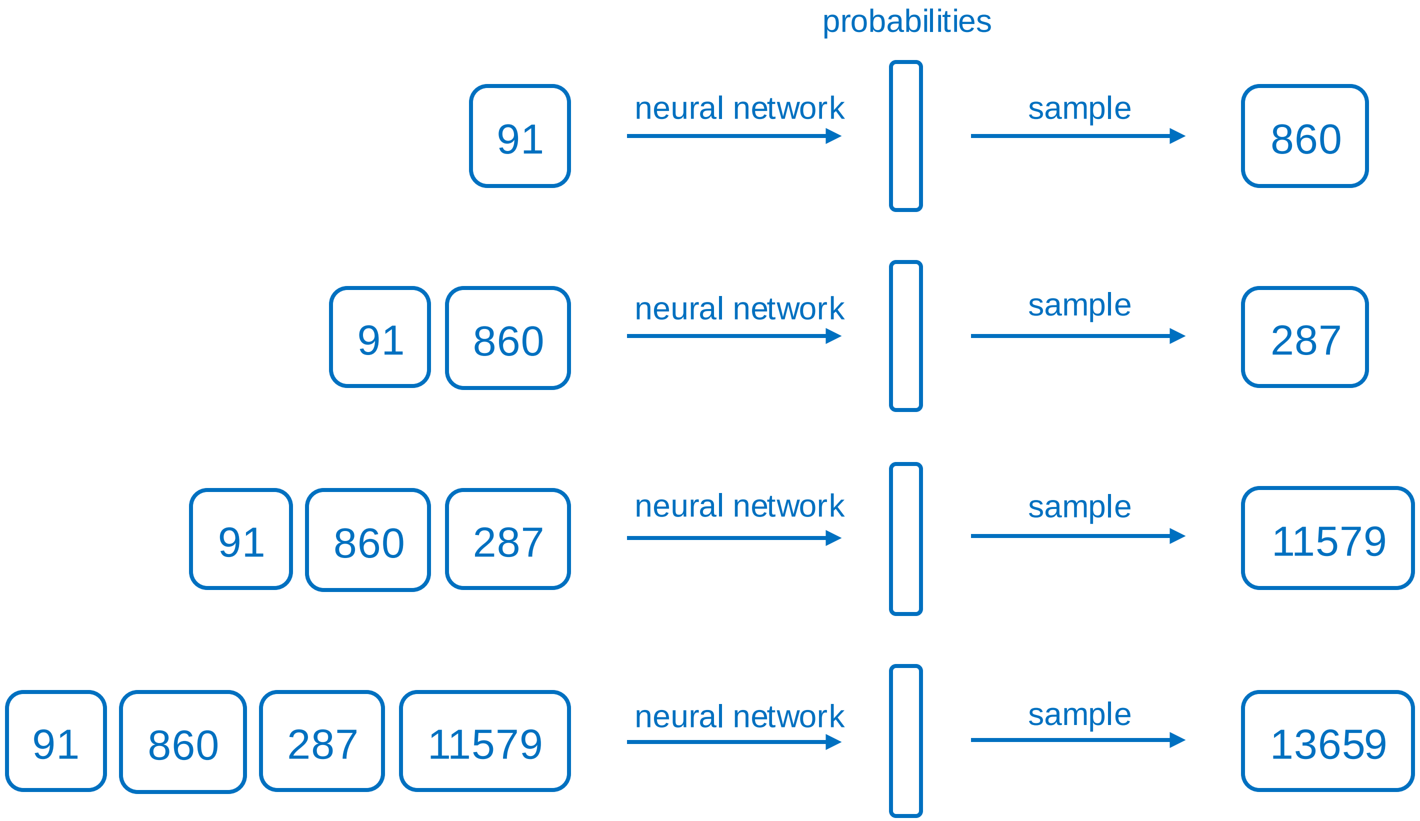
Neural network internals



- Parameters start randomly, resulting in random predictions
- Adjusting parameters (like tuning knobs on a DJ panel) makes predictions match patterns from training data

Inference

- To generate: Predict one token at a time



Example: GPT2 pretraining

- Train: 1558M parameters, 32000 steps, 1m tokens per step, 2.75s per step, ~24h
- Cost: 8xH100 gpu node, ~24h, $672 on cloud provider
- Buy and install GPU, expensive
- Luckily, we have open weights model

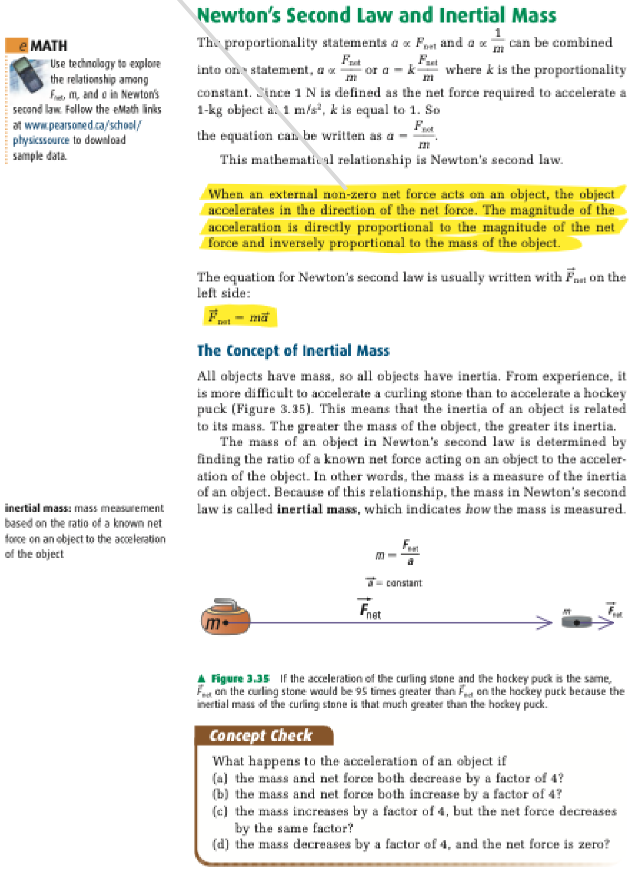
Demo of open weights model: Llama 3.1

- Autocomplete Wikipedia example: Zebra
- The knowledge cut-off of the model, example
- Sentence continuation example: what is 2+2?
- Need to make the model to be better at conversations

Overview: analogy of studying a textbook

Exposition ⇔ Pretraining
(background knowledge)

Worked problems ⇔ Supervised Finetuning
(problem + demonstrated solution, for imitation)



Newton's Second Law and Inertial Mass

e MATH
Use technology to explore the relationship among F_{net}, m, and a in Newton's second law. Follow the eMath links at www.pearsoned.ca/school/physicssource to download sample data.

The proportionality statements $a \propto F_{net}$ and $a \propto \frac{1}{m}$ can be combined into one statement, $a \propto \frac{F_{net}}{m}$ or $a = k\frac{F_{net}}{m}$ where k is the proportionality constant. Since 1 N is defined as the net force required to accelerate a 1-kg object at 1 m/s², k is equal to 1. So the equation can be written as $a = \frac{F_{net}}{m}$.

This mathematical relationship is Newton's second law.

When an external non-zero net force acts on an object, the object accelerates in the direction of the net force. The magnitude of the acceleration is directly proportional to the magnitude of the net force and inversely proportional to the mass of the object.

The equation for Newton's second law is usually written with $\vec{F}_{net}$ on the left side:

$$\vec{F}_{net} = m\vec{a}$$

The Concept of Inertial Mass

All objects have mass, so all objects have inertia. From experience, it is more difficult to accelerate a curling stone than to accelerate a hockey puck (Figure 3.35). This means that the inertia of an object is related to its mass. The greater the mass of the object, the greater its inertia.

The mass of an object in Newton's second law is determined by finding the ratio of a known net force acting on an object to the acceleration of the object. In other words, the mass is a measure of the inertia of an object. Because of this relationship, the mass in Newton's second law is called **inertial mass**, which indicates *how* the mass is measured.

inertial mass: mass measurement based on the ratio of a known net force on an object to the acceleration of the object

$$m = \frac{F_{net}}{a}$$

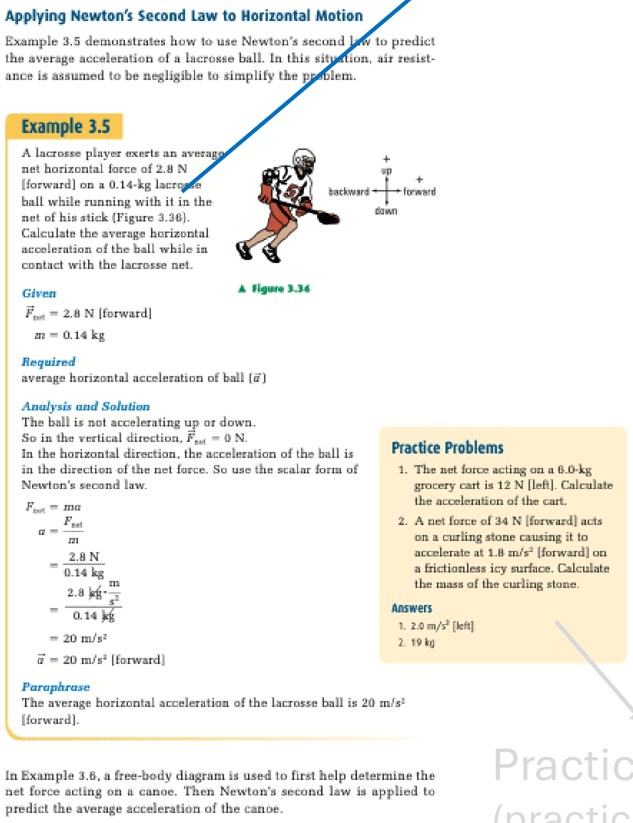
$\vec{a}$ = constant

$\vec{F}_{net}$

▲ **Figure 3.35** If the acceleration of the curling stone and the hockey puck is the same, $\vec{F}_{net}$ on the curling stone would be 95 times greater than $\vec{F}_{net}$ on the hockey puck because the inertial mass of the curling stone is that much greater than the hockey puck.

Concept Check

What happens to the acceleration of an object if
(a) the mass and net force both decrease by a factor of 4?
(b) the mass and net force both increase by a factor of 4?
(c) the mass increases by a factor of 4, but the net force decreases by the same factor?
(d) the mass decreases by a factor of 4, and the net force is zero?



Applying Newton's Second Law to Horizontal Motion

Example 3.5 demonstrates how to use Newton's second law to predict the average acceleration of a lacrosse ball. In this situation, air resistance is assumed to be negligible to simplify the problem.

Example 3.5

A lacrosse player exerts an average net horizontal force of 2.8 N [forward] on a 0.14-kg lacrosse ball while running with it in the net of his stick (Figure 3.36). Calculate the average horizontal acceleration of the ball while in contact with the lacrosse net.

backward, forward, up, down, +, +

▲ **Figure 3.36**

Given
$\vec{F}_{net}$ = 2.8 N [forward]
m = 0.14 kg

Required
average horizontal acceleration of ball ($\vec{a}$)

Analysis and Solution
The ball is not accelerating up or down.
So in the vertical direction, $\vec{F}_{net}$ = 0 N.
In the horizontal direction, the acceleration of the ball is in the direction of the net force. So use the scalar form of Newton's second law.

$$\begin{aligned} F_{net} &= ma \\ a &= \frac{F_{net}}{m} \\ &= \frac{2.8\ \text{N}}{0.14\ \text{kg}} \\ &= \frac{2.8\ \cancel{\text{kg}}\cdot\frac{\text{m}}{\text{s}^2}}{0.14\ \cancel{\text{kg}}} \\ &= 20\ \text{m/s}^2 \\ \vec{a} &= 20\ \text{m/s}^2\ [\text{forward}] \end{aligned}$$

Paraphrase
The average horizontal acceleration of the lacrosse ball is 20 m/s² [forward].

Practice Problems

1. The net force acting on a 6.0-kg grocery cart is 12 N [left]. Calculate the acceleration of the cart.
2. A net force of 34 N [forward] acts on a curling stone causing it to accelerate at 1.8 m/s² [forward] on a frictionless icy surface. Calculate the mass of the curling stone.

Answers
1. 2.0 m/s² [left]
2. 19 kg

In Example 3.6, a free-body diagram is used to first help determine the net force acting on a canoe. Then Newton's second law is applied to predict the average acceleration of the canoe.

Practice Problems ⇔ Reinforcement learning
(practice, trial & error until reach correct answer)

Supervised Finetuning (SFT): Teaching LLMs to Assist

- Goal: Train the model to respond like a helpful assistant.

- Input:
  - Base Model (trained on internet text, not yet conversational).
  - Labeled Conversations (human-written Q&A, multi-turn dialogues).
- Process:
  - Replace general internet data with conversation data.
  - Continue training so the model learns how assistants respond.
- Time Comparison: Months (pretraining) vs hours (SFT)
- Data Comparison: Trillions of tokens vs Millions of tokens

Conversation Data Format & Tokenization

- Tokenization Process
  - Uses the same method as pretraining, converting conversations into token sequences.
  - New Special Tokens added for better context.

- Real Datasets in Practice:
  - Labeling Instructions (e.g., InstructGPT Section 3.4).
  - Examples: OpenAssistant, UltraChat (with visualization).

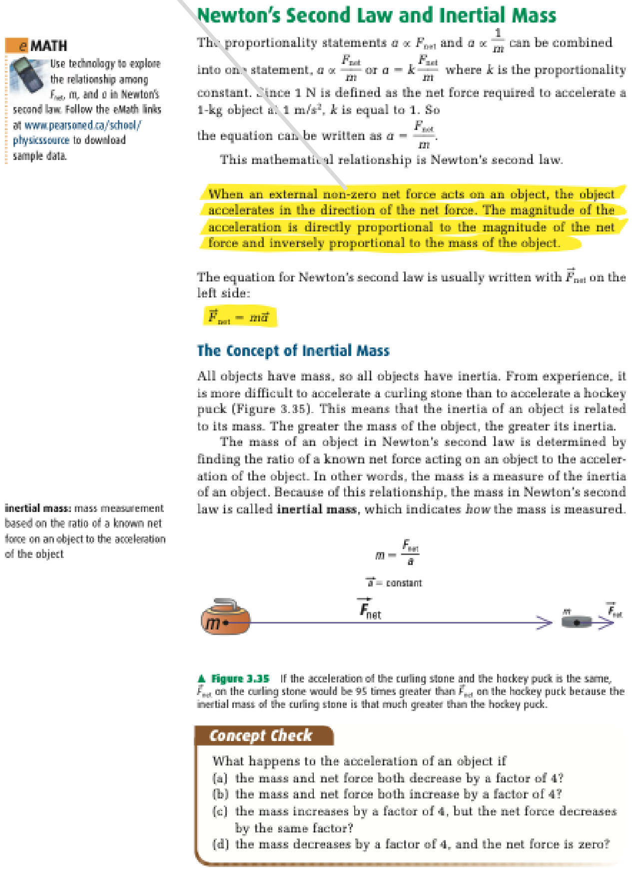
- Example:
  - Instruction-following model inference (e.g., Llama Instruct).

Conversation Data Format & Tokenization

- Tokenization Process
  - Uses the same method as pretraining, converting conversations into token sequences.
  - New Special Tokens added for better context.
- Real Datasets in Practice:
  - Labeling Instructions (e.g., InstructGPT Section 3.4).
  - Examples: OpenAssistant, UltraChat (with visualization).
- Example:
  - Instruction-following model inference (e.g., Llama Instruct).
- SFT model still need to generalized over the “example problems”

Overview: analogy of studying a textbook

Exposition ⇔ Pretraining
(background knowledge)


e MATH

Use technology to explore the relationship among F_{net}, m, and a in Newton's second law. Follow the eMath links at www.pearsoned.ca/school/physicssource to download sample data.

Newton's Second Law and Inertial Mass

The proportionality statements $a \propto F_{net}$ and $a \propto \frac{1}{m}$ can be combined into one statement, $a \propto \frac{F_{net}}{m}$ or $a = k\frac{F_{net}}{m}$ where k is the proportionality constant. Since 1 N is defined as the net force required to accelerate a 1-kg object at 1 m/s², k is equal to 1. So the equation can be written as $a = \frac{F_{net}}{m}$.

This mathematical relationship is Newton's second law.

When an external non-zero net force acts on an object, the object accelerates in the direction of the net force. The magnitude of the acceleration is directly proportional to the magnitude of the net force and inversely proportional to the mass of the object.

The equation for Newton's second law is usually written with $\vec{F}_{net}$ on the left side:

$$\vec{F}_{net} = m\vec{a}$$

The Concept of Inertial Mass

All objects have mass, so all objects have inertia. From experience, it is more difficult to accelerate a curling stone than to accelerate a hockey puck (Figure 3.35). This means that the inertia of an object is related to its mass. The greater the mass of the object, the greater its inertia.

The mass of an object in Newton's second law is determined by finding the ratio of a known net force acting on an object to the acceleration of the object. In other words, the mass is a measure of the inertia of an object. Because of this relationship, the mass in Newton's second law is called **inertial mass**, which indicates *how* the mass is measured.

inertial mass: mass measurement based on the ratio of a known net force on an object to the acceleration of the object

$$m = \frac{F_{net}}{a}$$

$\vec{a}$ = constant

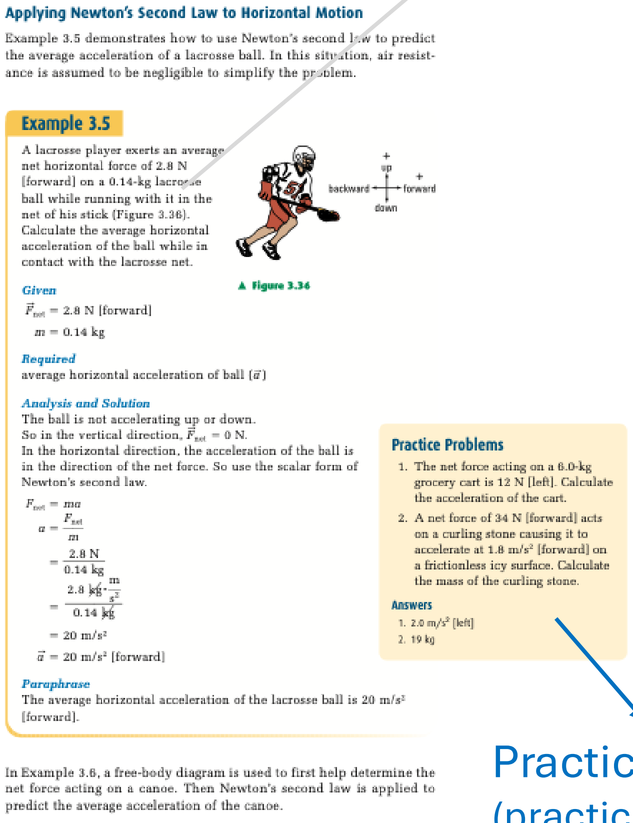
$\vec{F}_{net}$

▲ **Figure 3.35** If the acceleration of the curling stone and the hockey puck is the same, $\vec{F}_{net}$ on the curling stone would be 95 times greater than $\vec{F}_{net}$ on the hockey puck because the inertial mass of the curling stone is that much greater than the hockey puck.

Concept Check

What happens to the acceleration of an object if
(a) the mass and net force both decrease by a factor of 4?
(b) the mass and net force both increase by a factor of 4?
(c) the mass increases by a factor of 4, but the net force decreases by the same factor?
(d) the mass decreases by a factor of 4, and the net force is zero?

Worked problems ⇔ Supervised Finetuning
(problem + demonstrated solution, for imitation)


Applying Newton's Second Law to Horizontal Motion

Example 3.5 demonstrates how to use Newton's second law to predict the average acceleration of a lacrosse ball. In this situation, air resistance is assumed to be negligible to simplify the problem.

Example 3.5

A lacrosse player exerts an average net horizontal force of 2.8 N [forward] on a 0.14-kg lacrosse ball while running with it in the net of his stick (Figure 3.36). Calculate the average horizontal acceleration of the ball while in contact with the lacrosse net.

up, down, backward, forward

▲ **Figure 3.36**

Given
$\vec{F}_{net}$ = 2.8 N [forward]
m = 0.14 kg

Required
average horizontal acceleration of ball ($\vec{a}$)

Analysis and Solution
The ball is not accelerating up or down.
So in the vertical direction, $\vec{F}_{net}$ = 0 N.
In the horizontal direction, the acceleration of the ball is in the direction of the net force. So use the scalar form of Newton's second law.

$$\begin{aligned} F_{net} &= ma \\ a &= \frac{F_{net}}{m} \\ &= \frac{2.8 \text{ N}}{0.14 \text{ kg}} \\ &= \frac{2.8\ \cancel{\text{kg}} \cdot \frac{\text{m}}{\text{s}^2}}{0.14\ \cancel{\text{kg}}} \\ &= 20 \text{ m/s}^2 \\ \vec{a} &= 20 \text{ m/s}^2 \text{ [forward]} \end{aligned}$$

Paraphrase
The average horizontal acceleration of the lacrosse ball is 20 m/s² [forward].

Practice Problems

1. The net force acting on a 6.0-kg grocery cart is 12 N [left]. Calculate the acceleration of the cart.
2. A net force of 34 N [forward] acts on a curling stone causing it to accelerate at 1.8 m/s² [forward] on a frictionless icy surface. Calculate the mass of the curling stone.

Answers

1. 2.0 m/s² [left]
2. 19 kg

In Example 3.6, a free-body diagram is used to first help determine the net force acting on a canoe. Then Newton's second law is applied to predict the average acceleration of the canoe.

Practice Problems ⇔ Reinforcement learning
(practice, trial & error until reach correct answer)

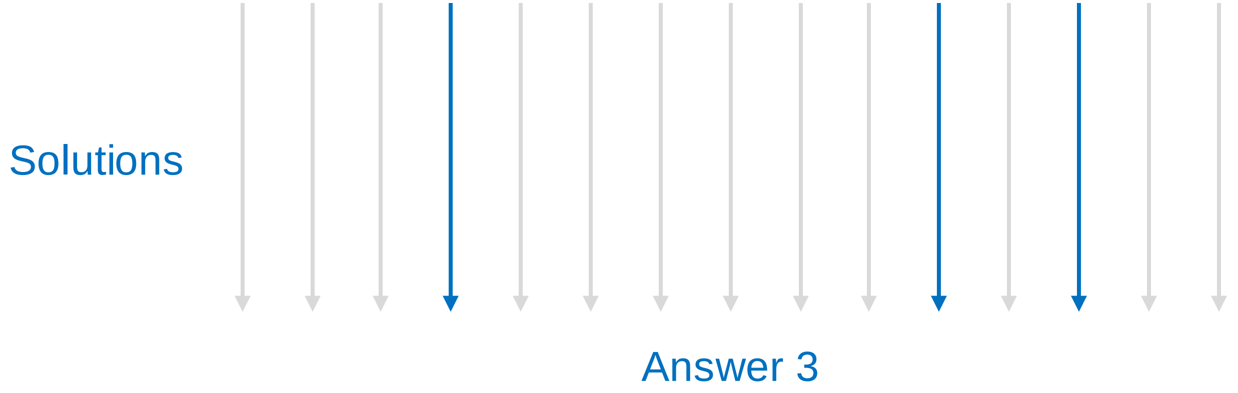
Reinforcement learning (RL)

- Goal: teach the model to figure out the solution itself

- Solution types:
  - verifiable solutions (e.g., math problems)
  - unverifiable solutions (e.g., jokes)

- RL for verifiable solutions
  - Generate multiple solutions
  - Take top solution
  - Train on it
  - Repeat (tens of thousands)

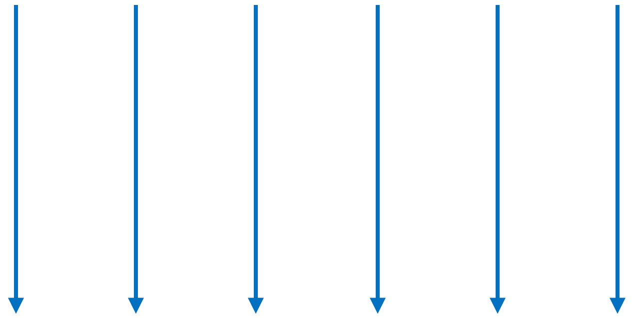
Emily buys 3 apples and 2 oranges. Each orange costs $2. The total cost of all the fruit is $13
What is the cost of each apple?



RL for Unverifiable Solutions

- Key Challenge: How to score generated outputs in subjective tasks?
- Example: Writing a joke—what makes it "good"?

“write a joke about Belgium weather”



Problem: how to score these automatically at scale?

Reward model scores:	0.1	0.8	0.3	0.4	0.5	0.7
Human ordering:	2	1	3	5	6	4

Naïve approach: 1,000 updates of 1,000 prompts of 1,000 runs (cost 1,000,000,000 scores from humans)

RLHF (human feedback) approach:

- Humans rank a few outputs instead of scoring each one.
- Train a reward model (NN simulator) based on preferences.
- Use the simulator to rank new outputs efficiently.

LLM Caveats

- Hallucination (demo)
- Shorter the response, less computation, worse results (demo)
- LLMs are not good at counting (demo)
- Other corner cases (demo)

References

- Andrej Karpathy, Deep Dive into LLMs like ChatGPT (link)

Outline

- Introduction to Large Language Models
- Training an LLM: Behind the Scenes
- Understanding the Costs
- Hands-on: Deploy Your Own Coding Copilot

Price Calculation

- GPU memory:
  - 4G VRAM per 1 billion float32 weights
  - 32b model: 128g VRAM -> 4x40G GPU or 6x24G GPU
  - There are also float16, float8, float4 mode
- Train
  - FLOPs per token: 6 x 32e9 = 1.92e11
  - Train on 1T tokens: 1.92e11 x 1e12 = 1.92e11 TFLOPs
  - On A100 312 TFLOPs: 1.92e11 / 312 = 6.15e9s = 237month
  - On 128 A100: 1.8 month, with overhead: 2.3 month
  - Price on cloud: 1.3USD/h *2.3*30*24*128 ~=275,558 USD
- Inference: one instance normally serve 10-20 users

Outline

- Introduction to Large Language Models
- Training an LLM: Behind the Scenes
- Understanding the Costs
- Hands-on: Deploy Your Own Coding Copilot

Code copilot

- Idea: complete, edit, question code
- Commercial products: Github copilot, cursor
- Local deploy:
  - Local model: HuggingFace, ollama
  - Code editor extension: continue.dev
- Demo:
  - Cursor
  - VSCode + continue.dev + ollama

Continue.dev

- Download and open Visual Studio Code
- Download continue extension
- Open config file of continue
- follow recommended config

Ollama

- Download install Ollama
- Pull models recommended by continue
- Check continue quick start

Thanks