Large Language Models

Table of contents

Machine learning, briefly 1
How LLMs actually work 1
Where it breaks: the apples 2
Why this matters for COMP1100 3

Machine learning, briefly

The seminar builds up from curve fitting. Given data on sleep hours against health, machine learning is the business of finding the function that fits:

- $y = ax + b$ — a straight line
- $y = ax^2 + b$ — a curve
- $y = ax + by + c$ — a plane, once there are two inputs (sleep *and* exercise)

The model is the function; learning is finding its parameters.

How LLMs actually work

An LLM predicts the **next token**, one at a time, from a probability distribution over the whole vocabulary.

Given the prefix *“The woman ran up the ____“*:

Candidate	Probability
“hill”	1.2%
“stairs”	1.9%
“the”	1.9%
“to”	1.2%

Candidate	Probability
“baking”	0.005%
“package”	0.0001%
“myriad”	0.000004%
“gazpacho”	0.00001%

Add one more word to the context — *“The woman ran up the”* — and the distribution shifts sharply: “hill” jumps to 10.8% and “stairs” to 6.9%, while “the” collapses to 0.00001%.

That’s the whole mechanism. And it explains why *“The woman ran up the ____“* and *“What did the woman run up? ____“* are, to the model, the same kind of problem: continue the text plausibly.

Where it breaks: the apples

The seminar’s worked failure, from ChatGPT:

> **Q:** I have 419 red apples and 201 green apples. I eat 300 of the apples. How many apples do I have left?
>
> **A:** 620 (initial count) − 300 (eaten) = **320 apples left.**

Correct. Then:

> **Q:** How many red apples do I have left?
>
> **A:** 419 − 300 = **119 red apples left.**
>
> **Q:** How many green apples do I have left?
>
> **A:** 201 − 300 = **−99 green apples.** It appears that you’ve eaten more green apples than you initially had…

The model subtracts the *whole* 300 from each colour independently, producing a negative count — and then narrates the absurdity rather than noticing it. The underlying problem was never specified (which apples were eaten?), and the model doesn’t flag the ambiguity; it generates a plausible continuation for each question in isolation.

The lesson for the course: an LLM produces text that *reads* like reasoning. Verify anything that has to be true.

Why this matters for COMP1100

The course's generative AI policy permits AI for anything **except generating interview or other data** (see comp1100[1]). The apples example is the argument for that rule — a model will happily produce fluent, confident, wrong content, and interview data that was generated rather than gathered defeats the entire point of customer-discovery[2].

See prompt-engineering[3] for getting useful work out of them anyway.

[1] courses/comp1100/comp1100.pdf
[2] courses/comp1100/customer-discovery.pdf
[3] courses/comp1100/prompt-engineering.pdf