The Path to AI

François Fleuret

Meta | UNIVERSITÉ DE GENÈVE

WIRED

SANDRO IANNACCONE SCIENCE DEC 24, 2025 5:00 AM

AlphaFold Changed Science. After 5 Years, It's Still Evolving

WIRED spoke with DeepMind's Pushmeet Kohli about the recent past—and promising future—of the Nobel Prize-winning research project that changed biology and chemistry forever.

AI at Meta @AIatMeta

To understand whether we're making genuine progress on reasoning, we entered our AI models in five STEM Olympiad competitions.

The results:
🏅 Asian Physics Olympiad (APhO): Perfect score, theory exam
🏅 International Physics Olympiad (IPhO): Perfect score, theory exam
🥇 International Mathematical Olympiad (IMO): Gold medal
🥇 International Chemistry Olympiad (IChO): Gold-medal-level performance
🥇 Romanian Masters of Mathematics (RMM): Gold-medal-level performance

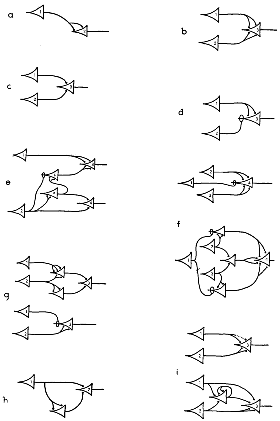
NewScientist

Mathematics

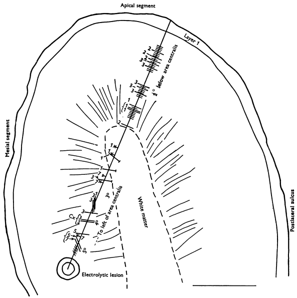
DeepMind and OpenAI claim gold in International Mathematical Olympiad

Two AI models have achieved gold medal standard for the first time in a prestigious competition for young mathematicians – and their developers claim these AIs could soon crack tough scientific problems

Cognition and learning can be engineered



(McCulloch and Pitts, 1943)



(Hubel and Wiesel, 1962)

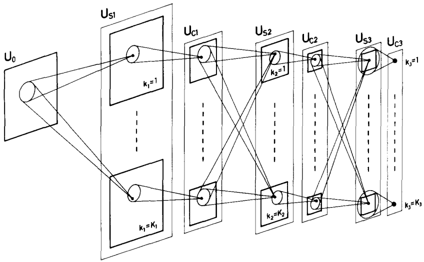
1949 – Donald Hebb proposes the Hebbian Learning principle (Hebb, 1949).

1951 – Marvin Minsky creates the first ANN (Hebbian learning, 40 neurons).

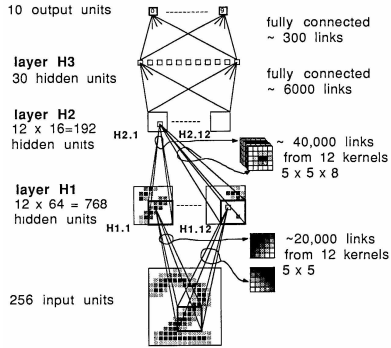
1958 – Frank Rosenblatt creates a perceptron to classify 20×20 images.

1959 – David H. Hubel and Torsten Wiesel demonstrate orientation selectivity and columnar organization in the cat's visual cortex (Hubel and Wiesel, 1962).

1982 – Paul Werbos proposes back-propagation for ANNs (Werbos, 1981).

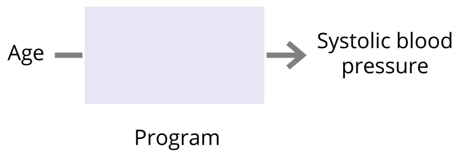


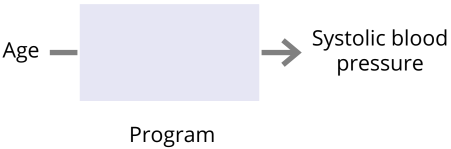
(Fukushima, 1980)

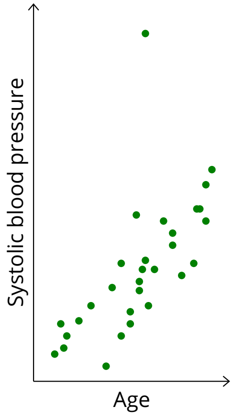


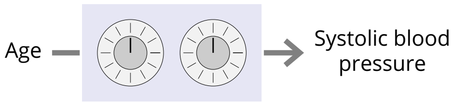
(LeCun et al., 1989)

Stupid learning leads to smart models

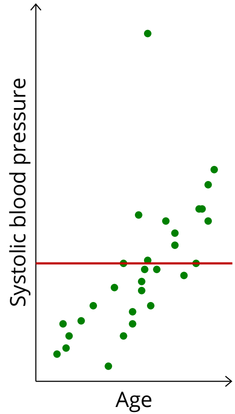

Age
Systolic blood pressure
Program

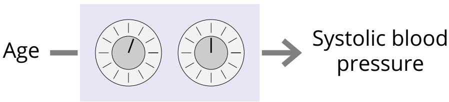

Age
Systolic blood pressure
Program


Systolic blood pressure
Age

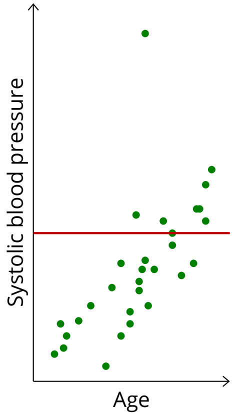


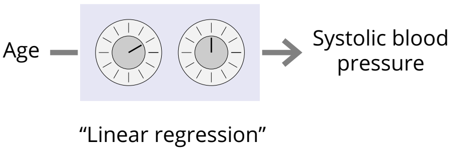
“Linear regression”

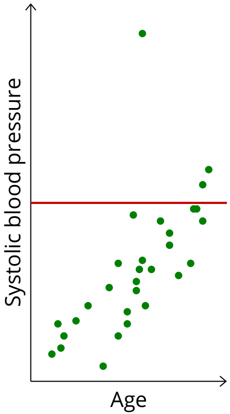


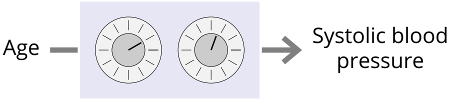


"Linear regression"

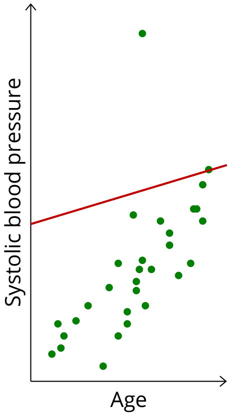


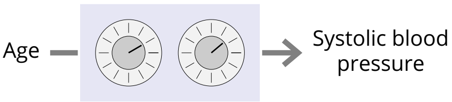

Age
Systolic blood pressure
"Linear regression"


Systolic blood pressure
Age

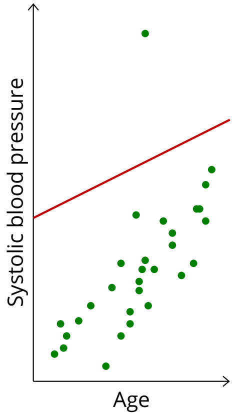


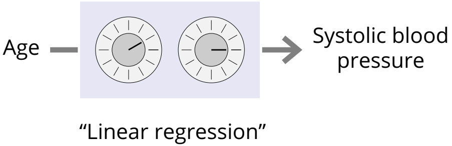
"Linear regression"

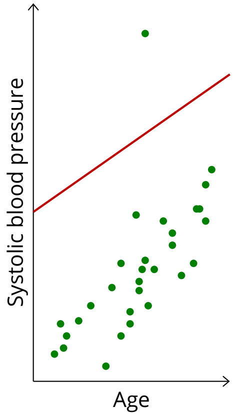


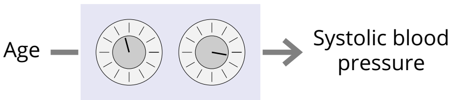


“Linear regression”

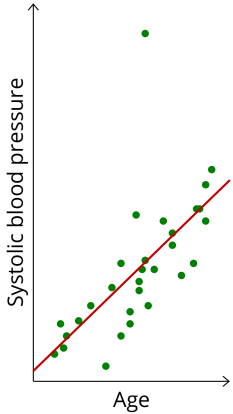



Age
Systolic blood pressure
“Linear regression”

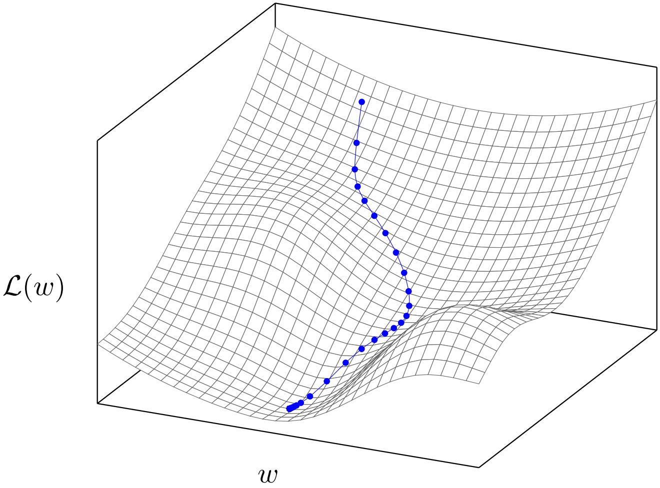

Systolic blood pressure
Age



“Linear regression”



In the general case, computing the proper parameters has to be done numerically through gradient descent.



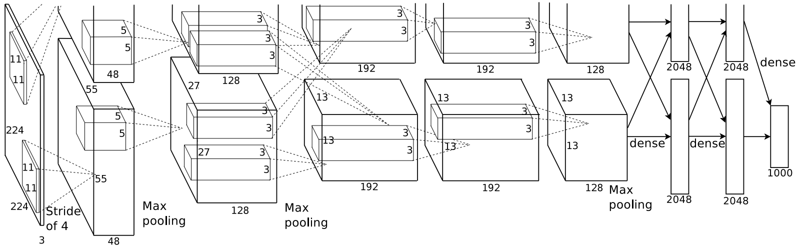
Model

```
model = nn.Sequential(
    nn.Conv2d( 1, 32, 5), nn.MaxPool2d(3), nn.ReLU(),
    nn.Conv2d(32, 64, 5), nn.MaxPool2d(2), nn.ReLU(),
    nn.Flatten(),
    nn.Linear(256, 200), nn.ReLU(),
    nn.Linear(200, 10)
)
```

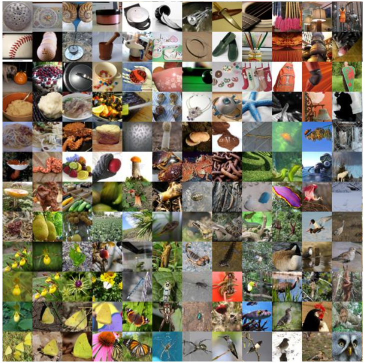
Training

```
criterion = nn.CrossEntropyLoss()

optimizer = torch.optim.SGD(model.parameters(), lr = 1e-2)

for e in range(nb_epochs):
    for input, target in data_loader_iterator(train_loader):
        output = model(input)
        loss = criterion(output, target)
        optimizer.zero_grad()
        loss.backward()
        optimizer.step()
```

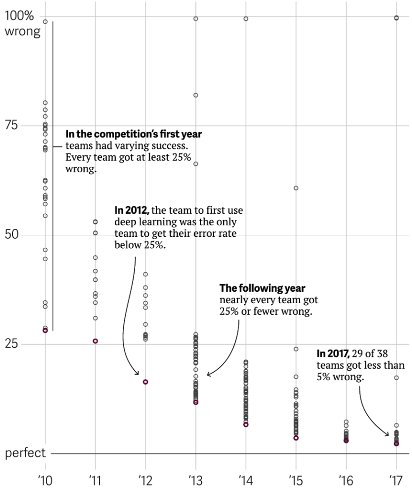
Quantity has a quality of its own



Very large models + GPUs (AlexNet, 2012)



ImageNet



(Gershgorn, 2017)

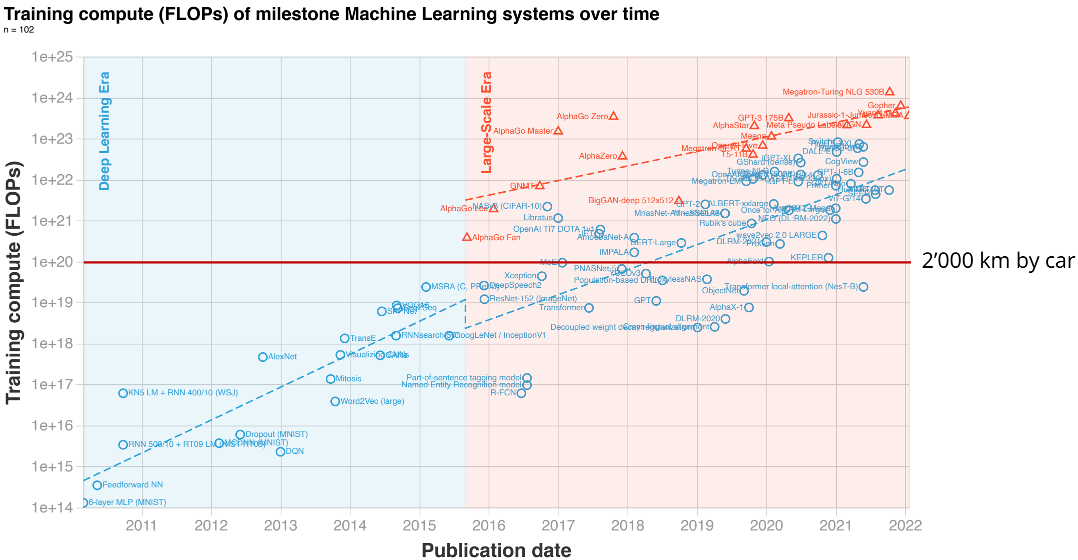
The last decade of progress in AI corresponds to a vast increase in the size of the “training sets”. The most successful deployed methods rely on human-labeled data.

Data-set	**Year**	**Nb. images**	**Size**
MNIST	1998	60K	12Mb
Caltech 256	2007	30K	1.2Gb
ImageNet	2012	1.2M	150Gb
LAION-5B	2022	5.85B	240Tb

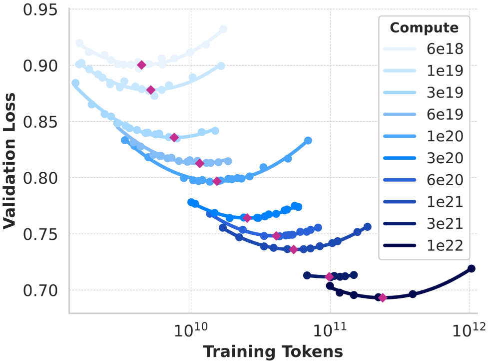
Data-set	**Year**	**Nb. books (250p)**	**Size***
SST2	2013	40	20Mb
WMT-18	2018	14K	7Gb
The Pile	2020	1.6M	825Gb
OSCAR	2020	12M	6Tb

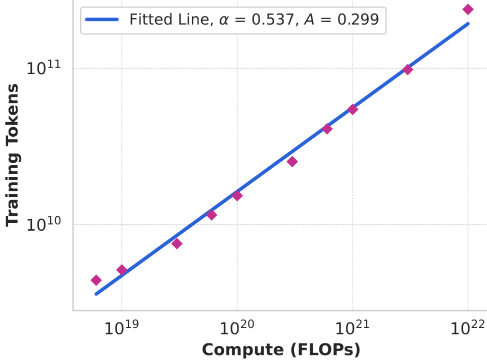
(* all the text of Wikipedia is 45Gb)

A $1'500 GPU executes $\simeq$ 35'000 billion operations per second.



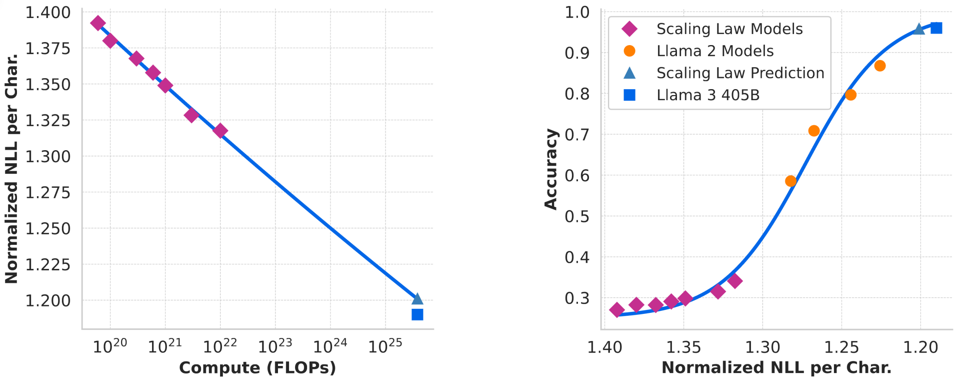
The performance of large models when the amount of compute and data set size increase follows very regular scaling laws.





(Dubey et al., 2024)

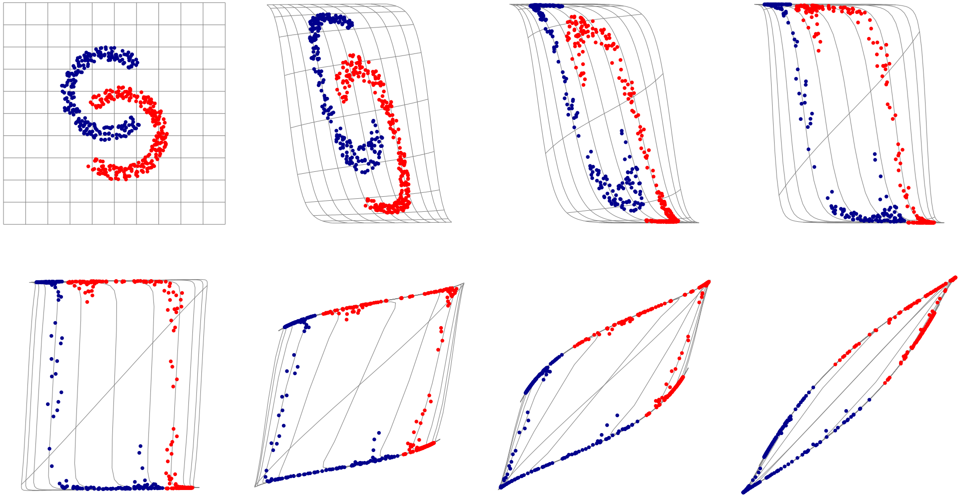
The performance of large models when the amount of compute and data set size increase follows very regular scaling laws.



(Dubey et al., 2024)

Optimizability trumps optimality

Using deeper architectures has been key in improving performance in many applications.



Four key elements allow the design of extremely deep architectures:

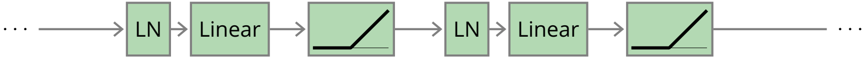
Four key elements allow the design of extremely deep architectures:

\+ Smart initialization to normalize all activations.

Four key elements allow the design of extremely deep architectures:

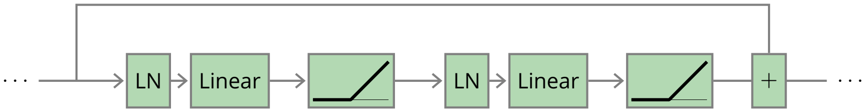
+ Smart initialization to normalize all activations.
+ Better non-linearities to prevent gradient vanishing.

Four key elements allow the design of extremely deep architectures:



+ Smart initialization to normalize all activations.
+ Better non-linearities to prevent gradient vanishing.
+ Normalization layers to keep the activation moments under control.

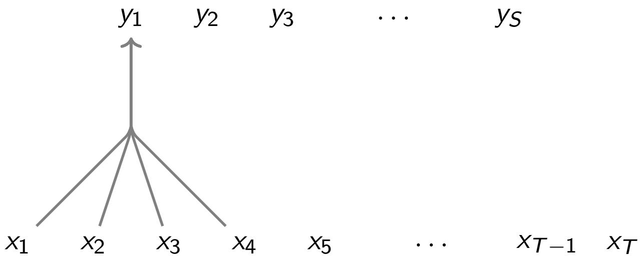
Four key elements allow the design of extremely deep architectures:



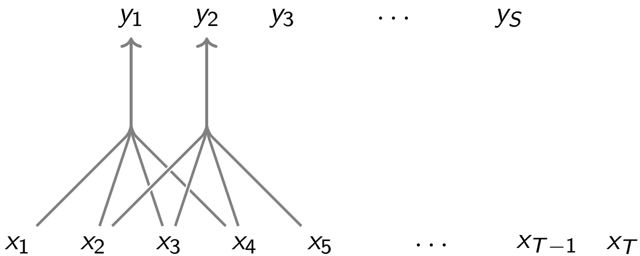
+ Smart initialization to normalize all activations.
+ Better non-linearities to prevent gradient vanishing.
+ Normalization layers to keep the activation moments under control.
+ Skip connections, to add the computed component to the current activations.

Attention is all you need

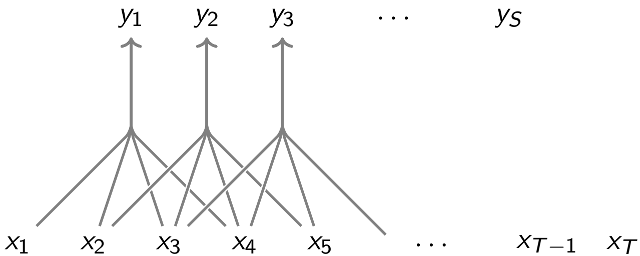
A convolution applies the **same operation** everywhere in the signal.



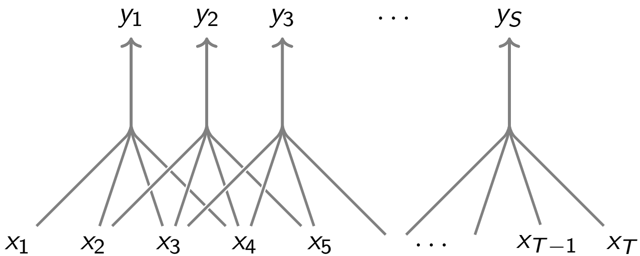
A convolution applies the **same operation** everywhere in the signal.



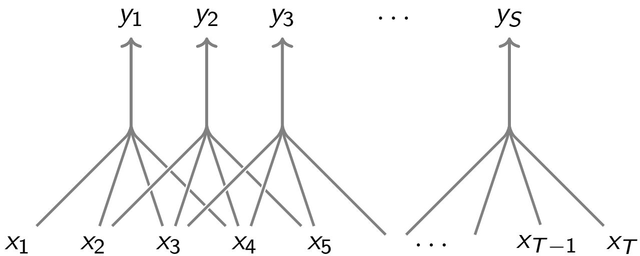
A convolution applies the **same operation** everywhere in the signal.



A convolution applies the **same operation** everywhere in the signal.



A convolution applies the **same operation** everywhere in the signal.



Such mechanisms are very efficient when

+ the signal is stationary, and
+ local structures are very informative.

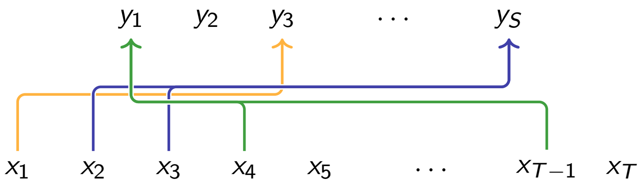
Some tasks involve more than local structures, e.g. translation:

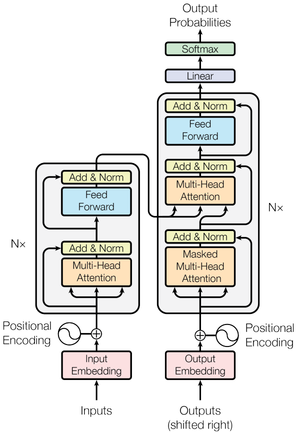
> “**An apple** that had been on the tree in the garden for weeks had finally been **picked up.**”
>
> “**Une pomme** qui était sur l’arbre du jardin depuis des semaines avait finalement été **ramassée.**”

Some tasks involve more than local structures, e.g. translation:

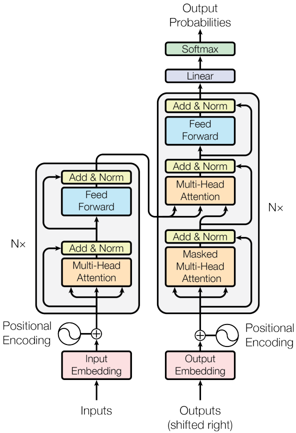
> “**An apple** that had been on the tree in the garden for weeks had finally been **picked up.**”
>
> “**Une pomme** qui était sur l’arbre du jardin depuis des semaines avait finalement été **ramassée.**”

It has motivated **attention-based processing** to transport information from parts of the signal to other parts dynamically identified.

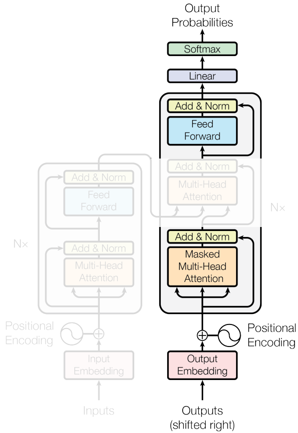




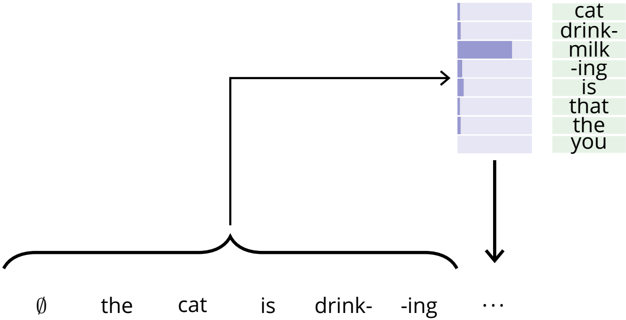
Transformer
(Vaswani et al., 2017)

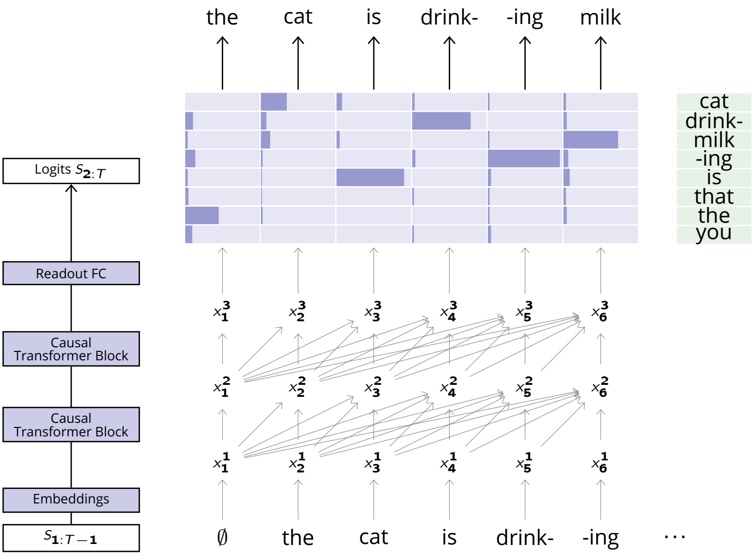


Transformer
(Vaswani et al., 2017)



Decoder-only Transformer
(Radford et al., 2018)


cat
drink-
milk
-ing
is
that
the
you
∅
the
cat
is
drink-
-ing
⋯



the
cat
is
drink-
-ing
milk
cat
drink-
milk
-ing
is
that
the
you
Logits $S_{2:T}$
Readout FC
Causal Transformer Block
Causal Transformer Block
Embeddings
$S_{1:T-1}$
x_1^3 x_2^3 x_3^3 x_4^3 x_5^3 x_6^3
x_1^2 x_2^2 x_3^2 x_4^2 x_5^2 x_6^2
x_1^1 x_2^1 x_3^1 x_4^1 x_5^1 x_6^1
∅
the
cat
is
drink-
-ing
…

Autoregression is smarter than expected

System Prompt (human-written)	*In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.*
Model Completion (machine-written, 10 tries)	`The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.` `Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.` `Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.`  `Pérez and the others then ventured further into the valley. "By the time we reached the top of one peak, the water looked blue, with some crystals on top," said Pérez.`  `Pérez and his friends were astonished to see the unicorn herd. These creatures could be seen from the air without having to move too much to see them – they were so close they could touch their horns.`

Such a “base model”, trained only to generate text, can already be put to use for instance for classification.

```
I: water boils at 100 degrees, O: physics.  I: the square root of two is irrational, O:
mathematics.  I: the set of prime numbers is infinite, O: mathematics.  I: gravity is
proportional to the mass, O:
```

Such a “base model”, trained only to generate text, can already be put to use for instance for classification.

```
I: water boils at 100 degrees, O: physics.  I: the square root of two is irrational, O:
mathematics.  I: the set of prime numbers is infinite, O: mathematics.  I: gravity is
proportional to the mass, O: physics.
```

Such a "base model", trained only to generate text, can already be put to use for instance for classification.

```
I: water boils at 100 degrees, O: physics.  I: the square root of two is irrational, O:
mathematics.  I: the set of prime numbers is infinite, O: mathematics.  I: gravity is
proportional to the mass, O: physics.
```

```
I: water boils at 100 degrees, O: physics.  I: the square root of two is irrational, O:
mathematics.  I: the set of prime numbers is infinite, O: mathematics.  I: squares are
rectangles, O:
```

Such a “base model”, trained only to generate text, can already be put to use for instance for classification.

```
I: water boils at 100 degrees, O: physics.  I: the square root of two is irrational, O: mathematics.  I: the set of prime numbers is infinite, O: mathematics.  I: gravity is proportional to the mass, O: physics.
```

```
I: water boils at 100 degrees, O: physics.  I: the square root of two is irrational, O: mathematics.  I: the set of prime numbers is infinite, O: mathematics.  I: squares are rectangles, O: mathematics.
```

Such a “base model”, trained only to generate text, can already be put to use for instance for classification.

```
I: water boils at 100 degrees, O: physics.  I: the square root of two is irrational, O:
mathematics.  I: the set of prime numbers is infinite, O: mathematics.  I: gravity is
proportional to the mass, O: physics.
```

```
I: water boils at 100 degrees, O: physics.  I: the square root of two is irrational, O:
mathematics.  I: the set of prime numbers is infinite, O: mathematics.  I: squares are
rectangles, O: mathematics.
```

```
I: I love apples, O: positive.  I: music is my passion, O: positive.  I: my job is
boring, O: negative.  I: frozen pizzas are awesome, O:
```

Such a “base model”, trained only to generate text, can already be put to use for instance for classification.

`I: water boils at 100 degrees, O: physics. I: the square root of two is irrational, O: mathematics. I: the set of prime numbers is infinite, O: mathematics. I: gravity is proportional to the mass, O: physics.`

`I: water boils at 100 degrees, O: physics. I: the square root of two is irrational, O: mathematics. I: the set of prime numbers is infinite, O: mathematics. I: squares are rectangles, O: mathematics.`

`I: I love apples, O: positive. I: music is my passion, O: positive. I: my job is boring, O: negative. I: frozen pizzas are awesome, O: positive.`

Such a “base model”, trained only to generate text, can already be put to use for instance for classification.

```
I: water boils at 100 degrees, O: physics.  I: the square root of two is irrational, O:
mathematics.  I: the set of prime numbers is infinite, O: mathematics.  I: gravity is
proportional to the mass, O: physics.
```

```
I: water boils at 100 degrees, O: physics.  I: the square root of two is irrational, O:
mathematics.  I: the set of prime numbers is infinite, O: mathematics.  I: squares are
rectangles, O: mathematics.
```

```
I: I love apples, O: positive.  I: music is my passion, O: positive.  I: my job is
boring, O: negative.  I: frozen pizzas are awesome, O: positive.
```

```
I: I love apples, O: positive.  I: music is my passion, O: positive.  I: my job is
boring, O: negative.  I: frozen pizzas taste like cardboard, O:
```

Such a “base model”, trained only to generate text, can already be put to use for instance for classification.

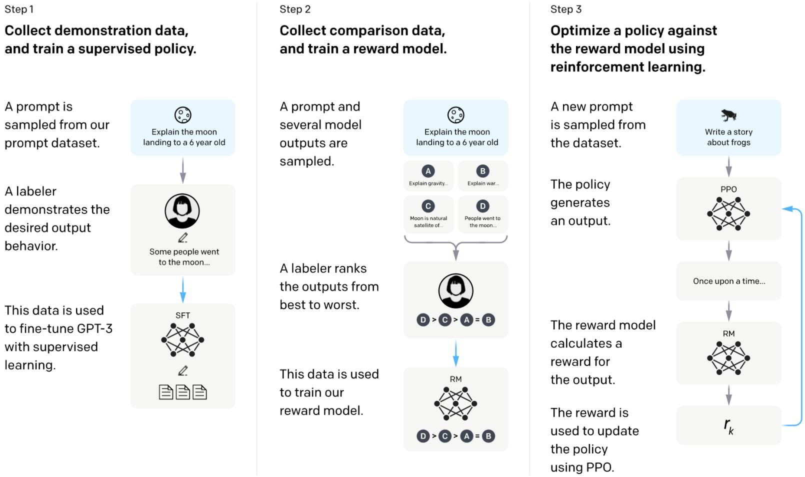
`I: water boils at 100 degrees, O: physics. I: the square root of two is irrational, O: mathematics. I: the set of prime numbers is infinite, O: mathematics. I: gravity is proportional to the mass, O: physics.`

`I: water boils at 100 degrees, O: physics. I: the square root of two is irrational, O: mathematics. I: the set of prime numbers is infinite, O: mathematics. I: squares are rectangles, O: mathematics.`

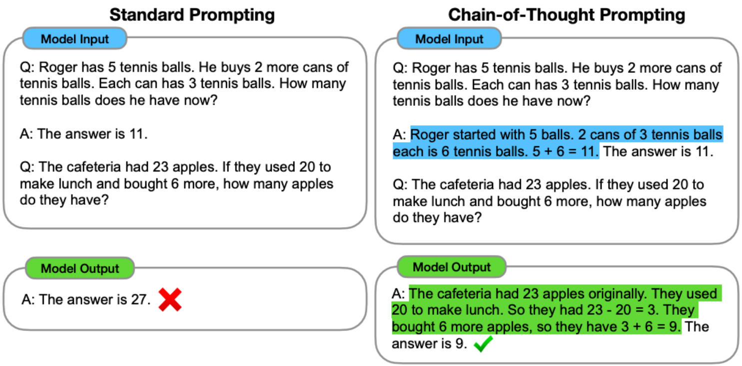
`I: I love apples, O: positive. I: music is my passion, O: positive. I: my job is boring, O: negative. I: frozen pizzas are awesome, O: positive.`

`I: I love apples, O: positive. I: music is my passion, O: positive. I: my job is boring, O: negative. I: frozen pizzas taste like cardboard, O: negative.`

For ChatGPT, for instance, base training is followed by "Reinforcement Learning from Human Feedback" (RLHF) to create an assistant.



It was then discovered that prompting the model with "step-by-step" reasoning examples dramatically improved its performance.



This can be scaled up to complex reasoning, e.g. mathematics, with reinforcement learning to boost the generation of chains of thought.

Conclusion

+ Large language models are extraordinary general-purpose “cognitive engines” well suited to modern computational devices.

+ Large language models are extraordinary general-purpose “cognitive engines” well suited to modern computational devices.
+ The current trend focuses on improving training procedures to leverage text generation for reasoning and agentic behavior.

+ Large language models are extraordinary general-purpose “cognitive engines” well suited to modern computational devices.
+ The current trend focuses on improving training procedures to leverage text generation for reasoning and agentic behavior.
+ We have probably passed the “data peak” and need new ways to keep improving. Synthetic data and self play are probably key.

+ Large language models are extraordinary general-purpose "cognitive engines" well suited to modern computational devices.
+ The current trend focuses on improving training procedures to leverage text generation for reasoning and agentic behavior.
+ We have probably passed the "data peak" and need new ways to keep improving. Synthetic data and self play are probably key.
+ Architectures are lacking fundamental capabilities such as continuous and multi-modal reasoning.

Questions?

A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, et al. **The llama 3 herd of models**, 2024.

K. Fukushima. **Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position**. *Biological Cybernetics*, 36(4):193–202, April 1980.

D. Gershgorn. **The data that transformed AI research—and possibly the world**, July 2017.

D. O. Hebb. *The organization of behavior: A neuropsychological theory*. Wiley, 1949. ISBN 0-8058-4300-0.

D. Hubel and T. Wiesel. **Receptive fields, binocular interaction, and functional architecture in the cat's visual cortex**. *Journal of Physiology*, 160:106–154, 1962.

Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. **Backpropagation applied to handwritten zip code recognition**. *Neural Computation*, 1(4): 541–551, 1989.

W. S. McCulloch and W. Pitts. **A logical calculus of the ideas immanent in nervous activity**. *The bulletin of mathematical biophysics*, 5(4):115–133, 1943.

A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever. **Improving language understanding by generative pre-training**, 2018.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin. **Attention is all you need**. *CoRR*, abs/1706.03762, 2017.

P. J. Werbos. **Applications of advances in nonlinear sensitivity analysis**. In *Proceedings of the 10th IFIP Conference*, pages 762–770, 1981.