Training and deploying LLMs reliably

The good, the bad and ugly, and the useful!

Arjun Chandra
Founder brua.io

21 Feb 2023, CuttingEdgeAI, Domus Bibliotheca, Oslo

@boelger

Quick intro...

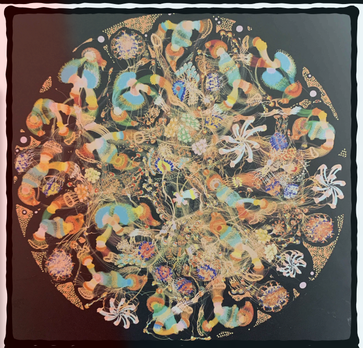
Father of **two amazing humans** and a few quirky artificial minds

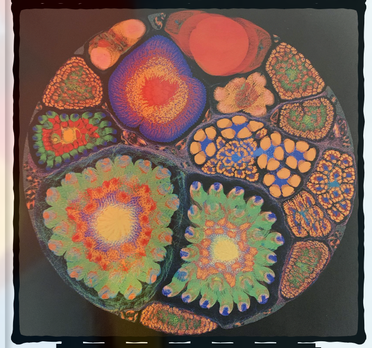














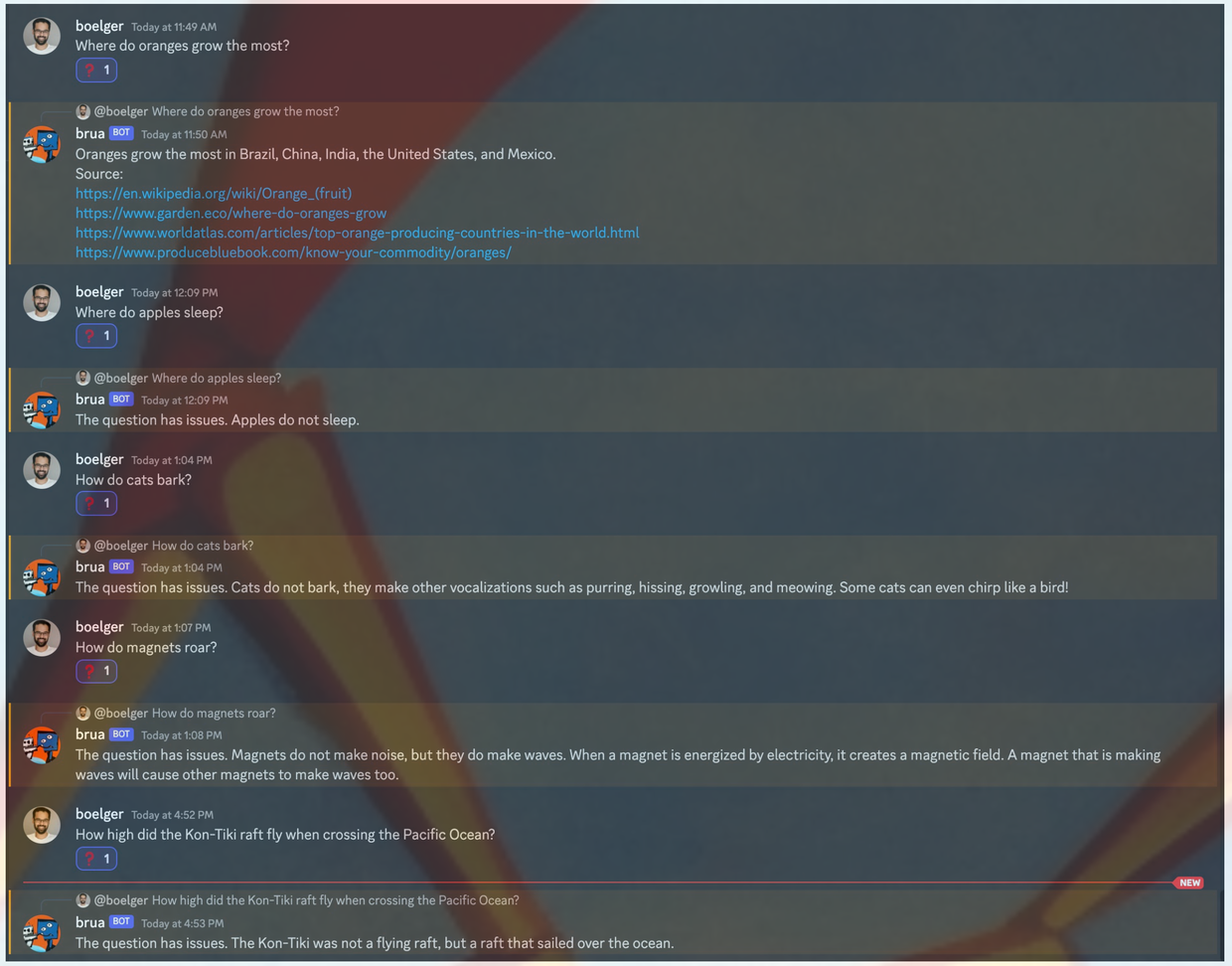
Artificial minds side more formally...

- **Current work**: Building brua.io — enabling the use of AI capabilities to **augment human endeavours reliably and safely**
- **Recent work**: Lead the **large model training stability program** at Graphcore — **stable training** of **large (foundation) models** in **reduced precision**
- **Background**
  - Academia (1999 - 2013): from home made **robots** to **computational self-awareness**
  - Entrepreneurship (2012 - 2015): **AI driven** personalised math **education**
  - Industry (2016 - 2022): surfing the **deep learning wave** and **building technology, competence, communities, and product**



Brua the Critical

Composition of LLMs and an expanding suite of reliability capabilities

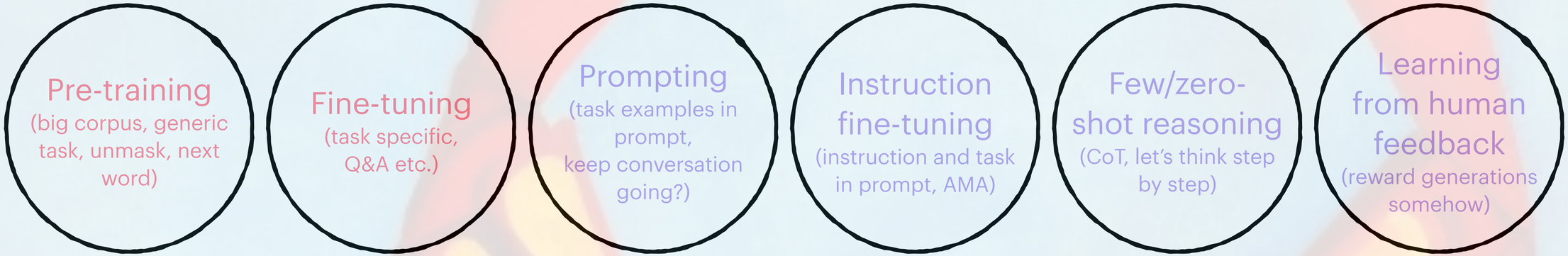


@boelger

The good

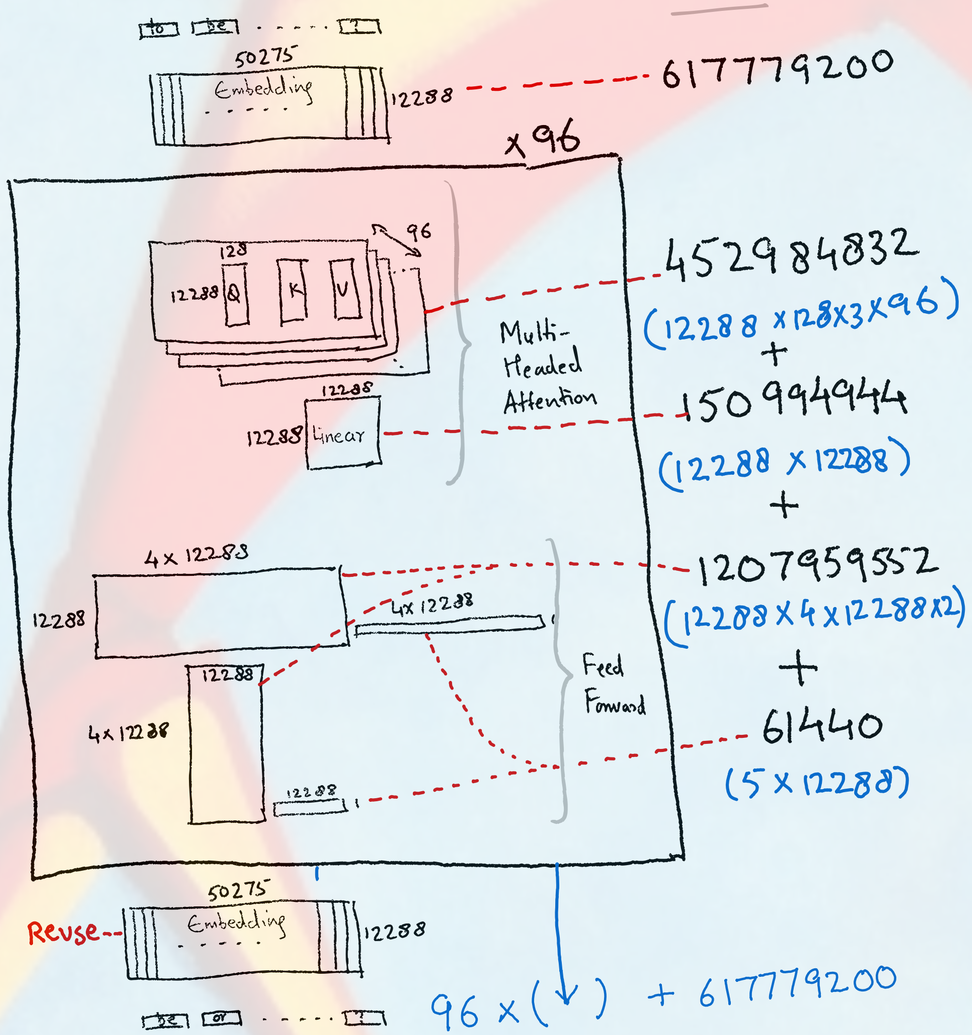
LLM training paradigms in flux

Capabilities emerging



@boelger

GPT-3



$96 \times (\downarrow) + 617792000$

$96 \times 1812000768 + 617792000$

$\downarrow$

174569852928

~ 175 billion

@boelger

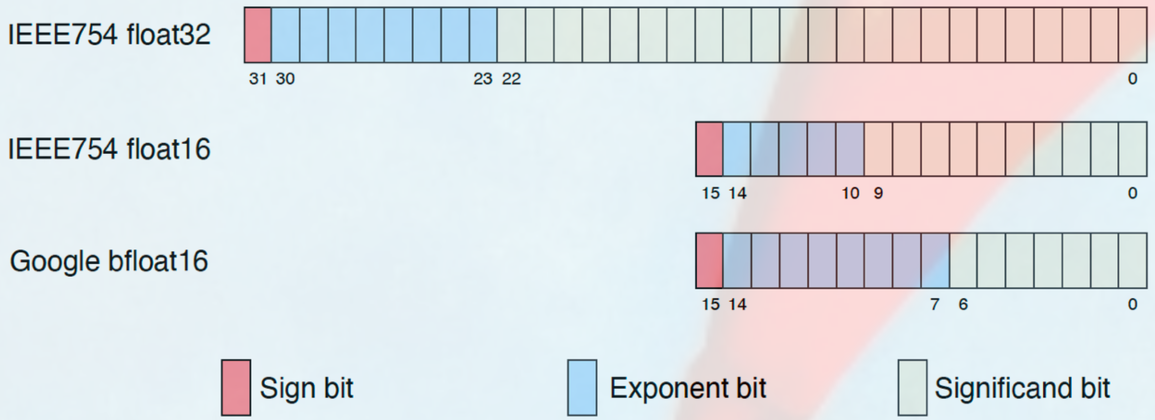
Stable training of large (foundation) models in reduced precision on IPU systems

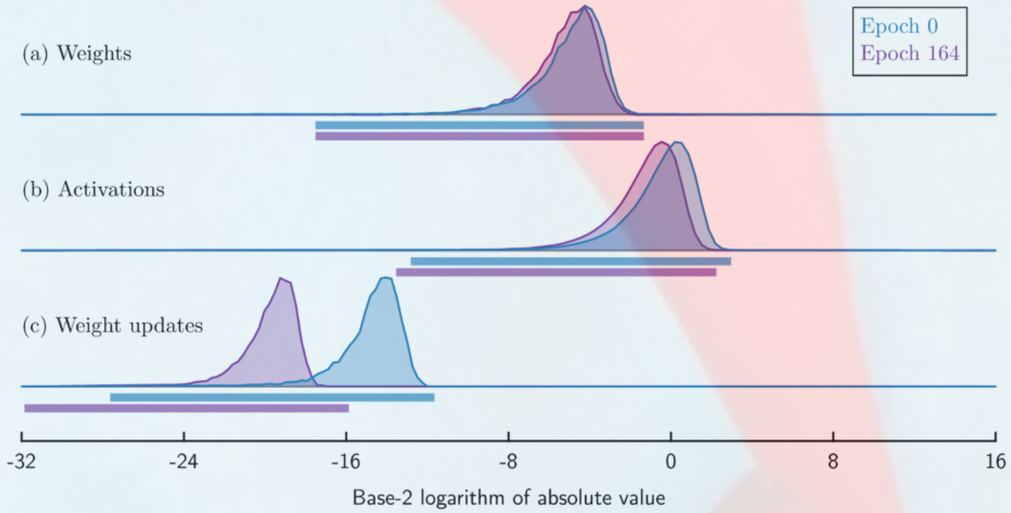
PyTorch	Enable Automatic Loss Scaling (ALS) for PyTorch model wrapped in a PopTorch **training wrapper** `opts.Training.setAutomaticLossScaling(True)`
	↓ ONNX++
PopART	Manipulate IR via ALS and helper transformations for **closed loop control** over gradient scales via statistics derived from the gradients
	↓ IR
Poplar	Expose efficient implementation of Ops to PopART, e.g. **gather tensor statistics**, and compile optimised graph targeting IPU device

↓ Compiled Poplar Graph



Why reduce precision?





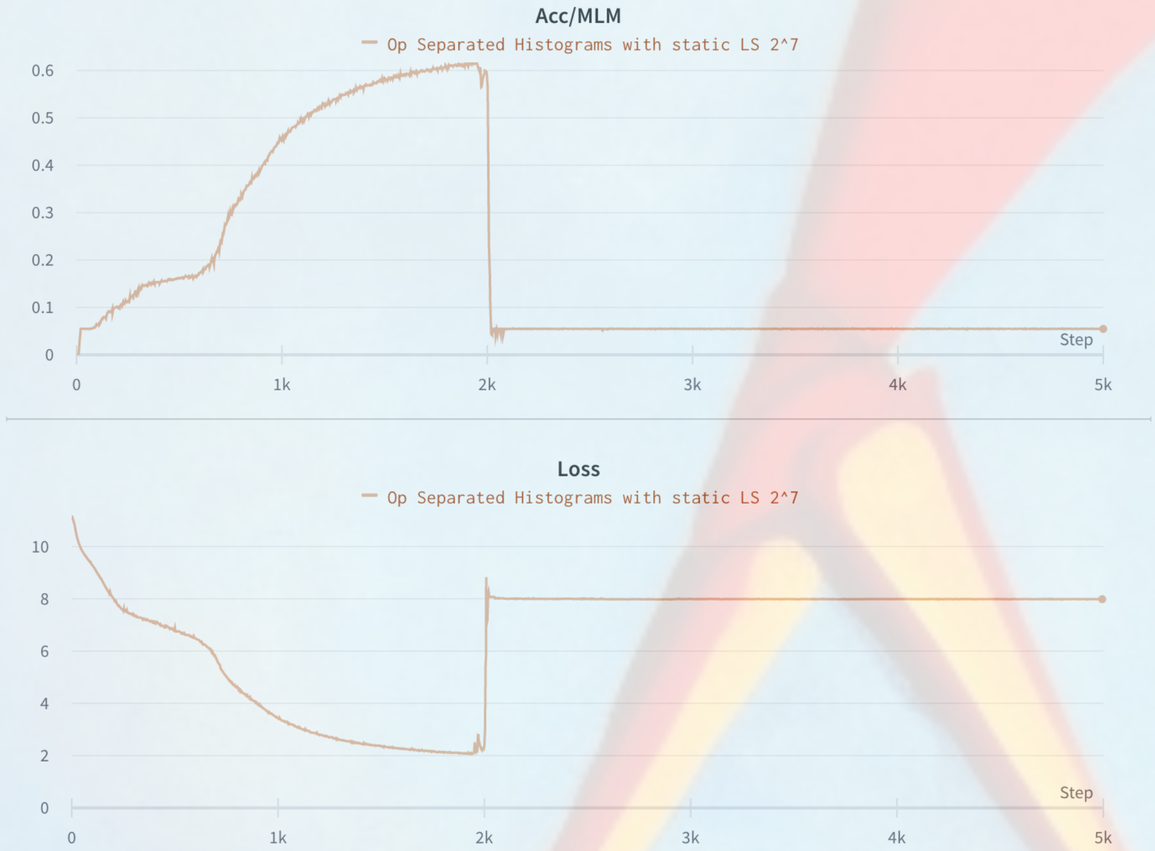
- Large compute, memory, bandwidth, and thus a large energy footprint [1]
- Reducing precision (e.g. halving) can alleviate these issues and **accelerate training** and **unblock experimentation**, e.g. multiplies scale quadratically with #bits
- May lead to **instabilities** due to underflow/overflow

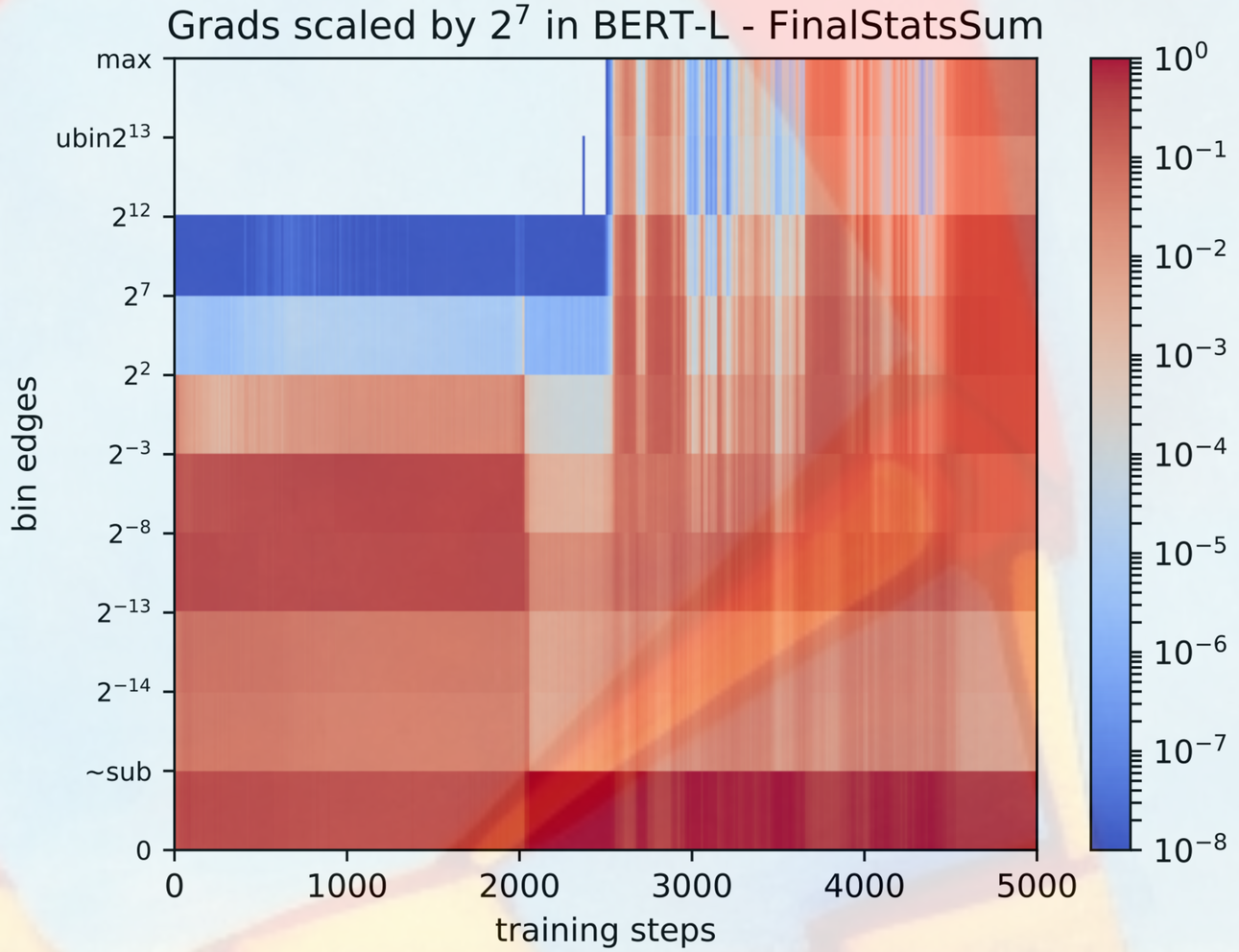
[1] Luccioni et al. (2022), **Estimating the Carbon Footprint of BLOOM, a 176B Parameter Language Model**, https://arxiv.org/abs/2211.02001

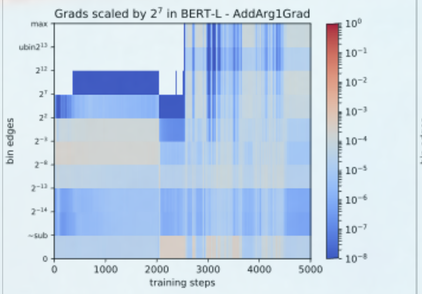
[2] Image on right adapted from Köster et al. (2017) **Flexpoint: An Adaptive Numerical Format for Efficient Training of Deep Neural Networks,** https://arxiv.org/abs/1711.02213

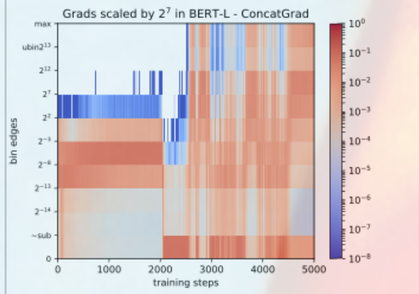
Large model training is numerically unstable in reduced precision

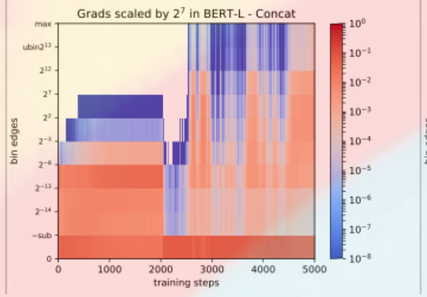
Failure mode: late detrimental underflow

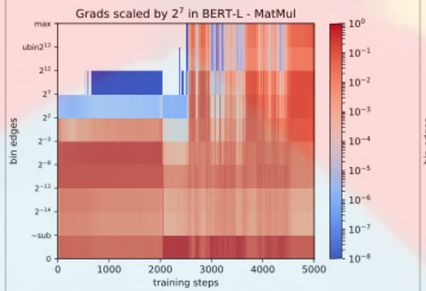


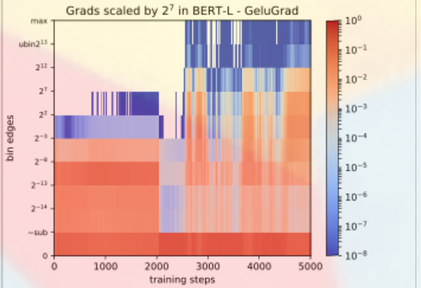


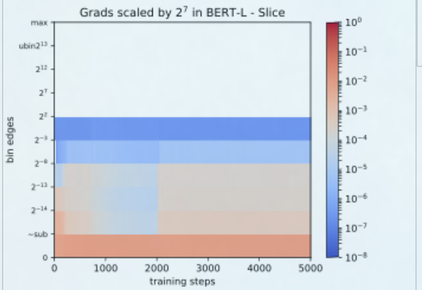

Grads scaled by 2^7 in BERT-L - AddArg1Grad
bin edges
training steps

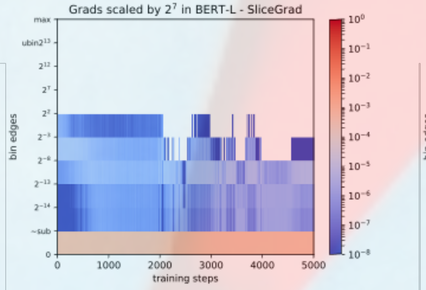

Grads scaled by 2^7 in BERT-L - ConcatGrad
bin edges
training steps

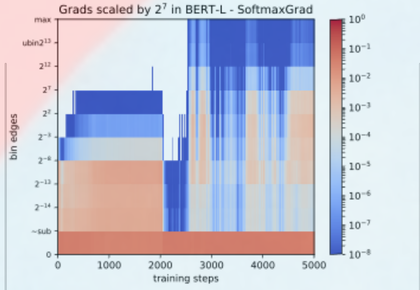

Grads scaled by 2^7 in BERT-L - Concat
bin edges
training steps

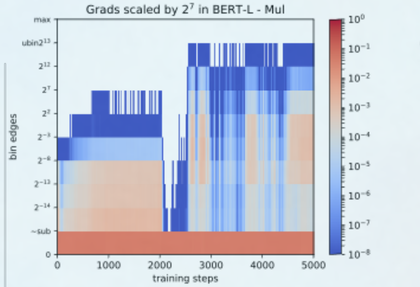

Grads scaled by 2^7 in BERT-L - MatMul
bin edges
training steps

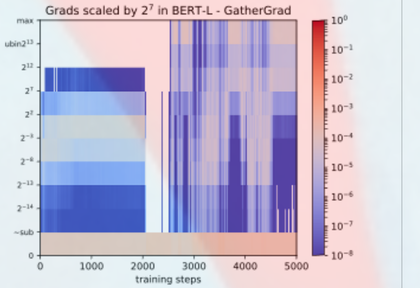

Grads scaled by 2^7 in BERT-L - GeluGrad
bin edges
training steps

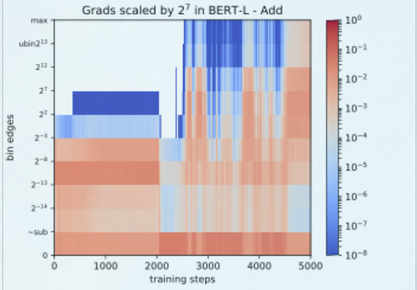

Grads scaled by 2^7 in BERT-L - Slice
bin edges
training steps

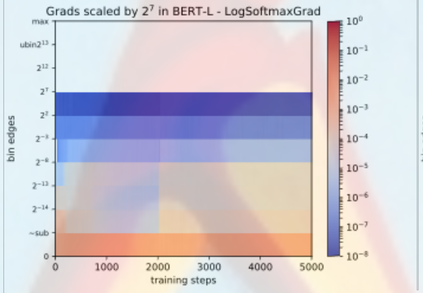

Grads scaled by 2^7 in BERT-L - SliceGrad
bin edges
training steps

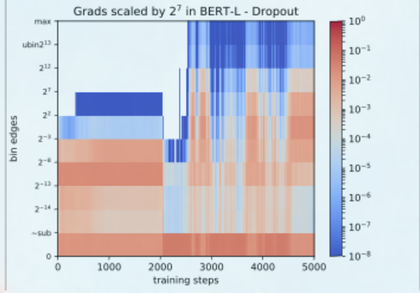

Grads scaled by 2^7 in BERT-L - SoftmaxGrad
bin edges
training steps

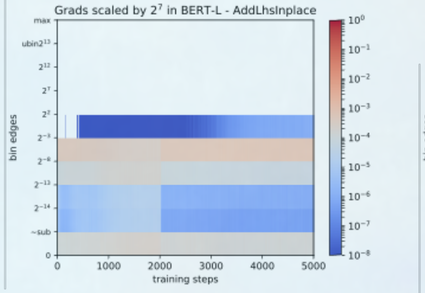

Grads scaled by 2^7 in BERT-L - Mul
bin edges
training steps

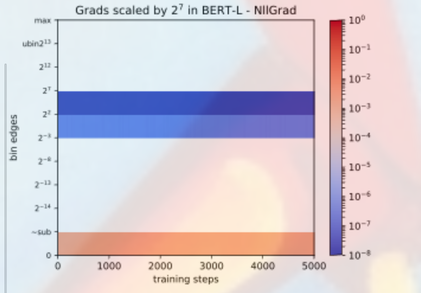

Grads scaled by 2^7 in BERT-L - GatherGrad
bin edges
training steps

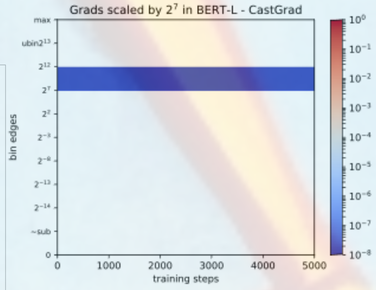

Grads scaled by 2^7 in BERT-L - Add
bin edges
training steps

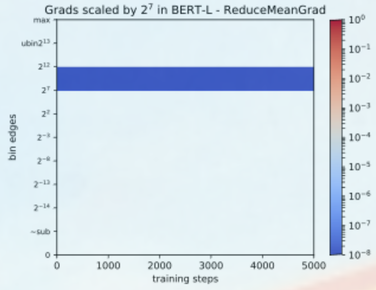

Grads scaled by 2^7 in BERT-L - LogSoftmaxGrad
bin edges
training steps


Grads scaled by 2^7 in BERT-L - Dropout
bin edges
training steps


Grads scaled by 2^7 in BERT-L - AddLhsInplace
bin edges
training steps

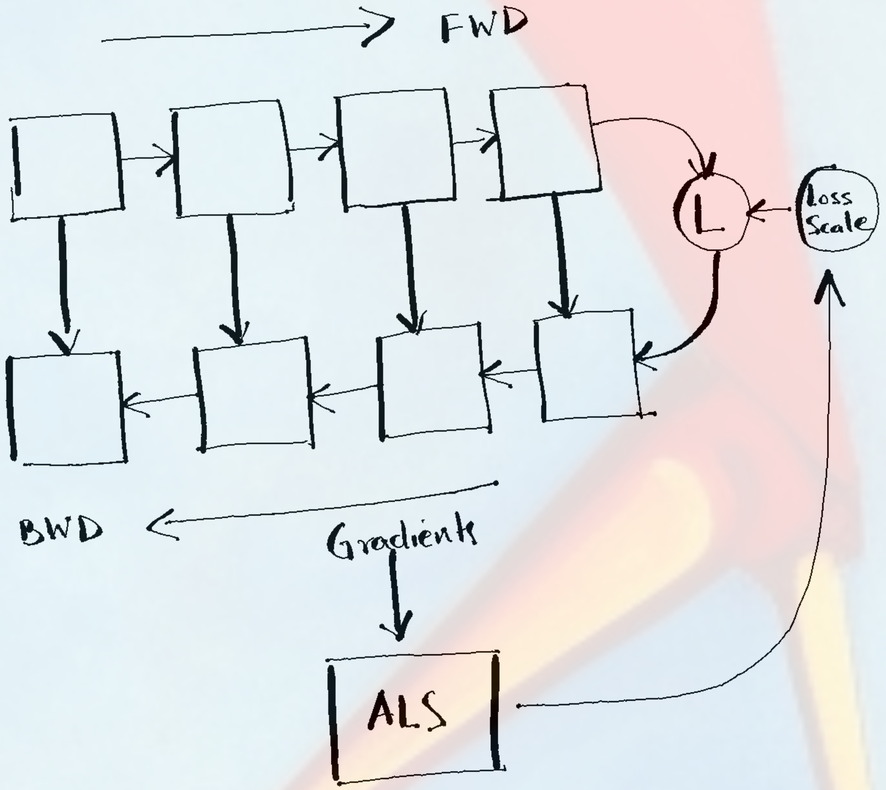

Grads scaled by 2^7 in BERT-L - NllGrad
bin edges
training steps


Grads scaled by 2^7 in BERT-L - CastGrad
bin edges
training steps


Grads scaled by 2^7 in BERT-L - ReduceMeanGrad
bin edges
training steps

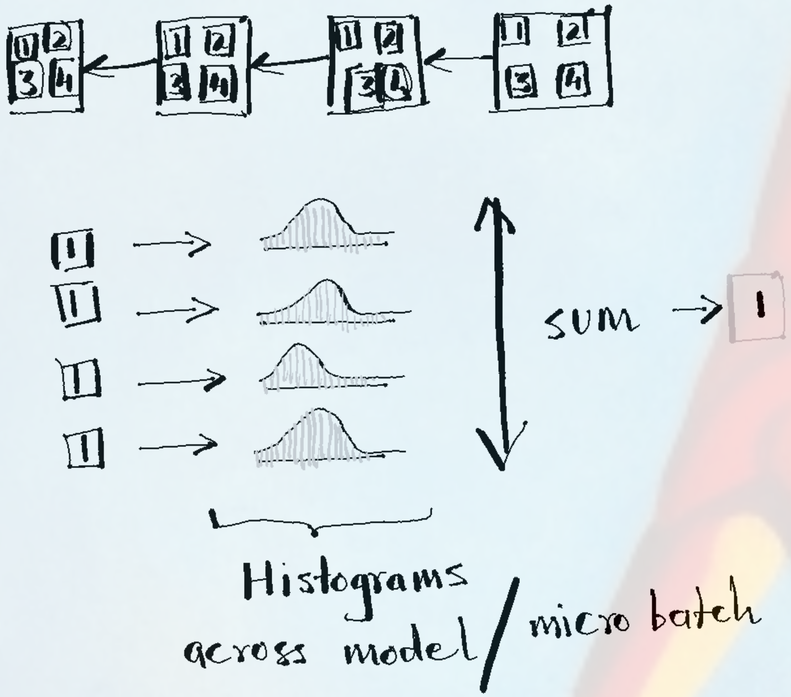
Automatic loss scaling for stable training of large models

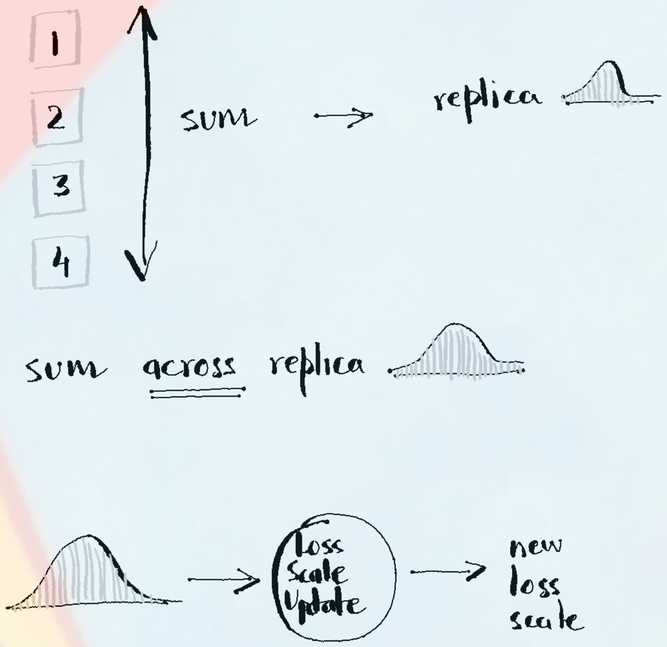
- Closed loop control
- Gradients observed and loss scale updated every optimiser step
- Histograms
- FP32 loss scale to deal with underflow
- Mixed precision math support

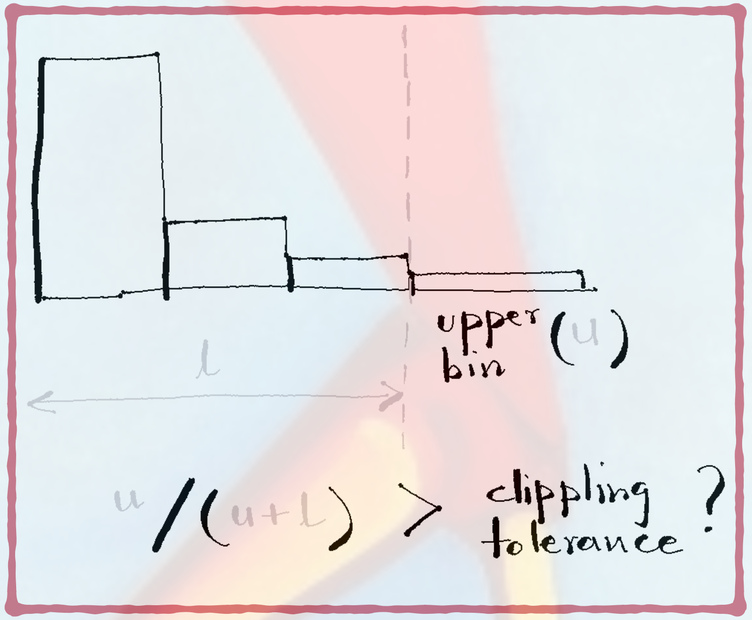


Reduced statistic used to make loss scale update decision

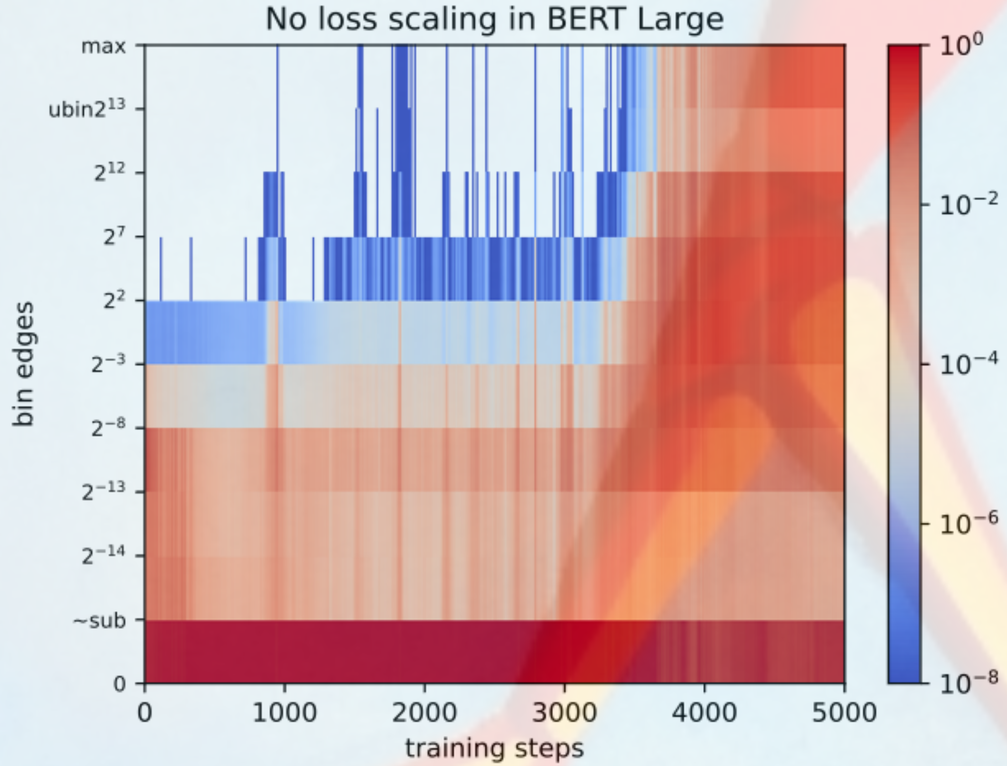
Gradient histograms are summed across **Ops**, across **micro-batches**, and across **replicas**

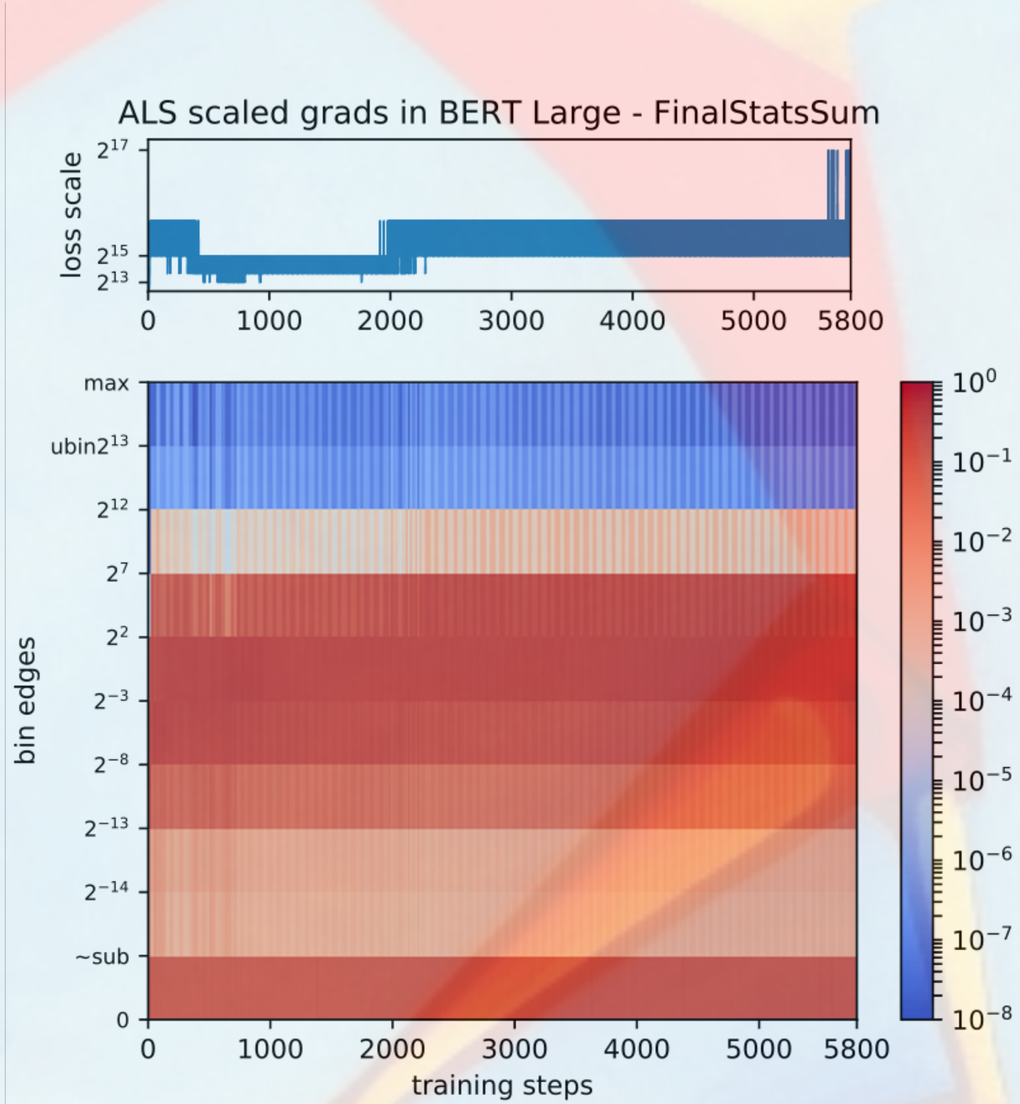






Gradients scaled at different phases of training appropriately with ALS

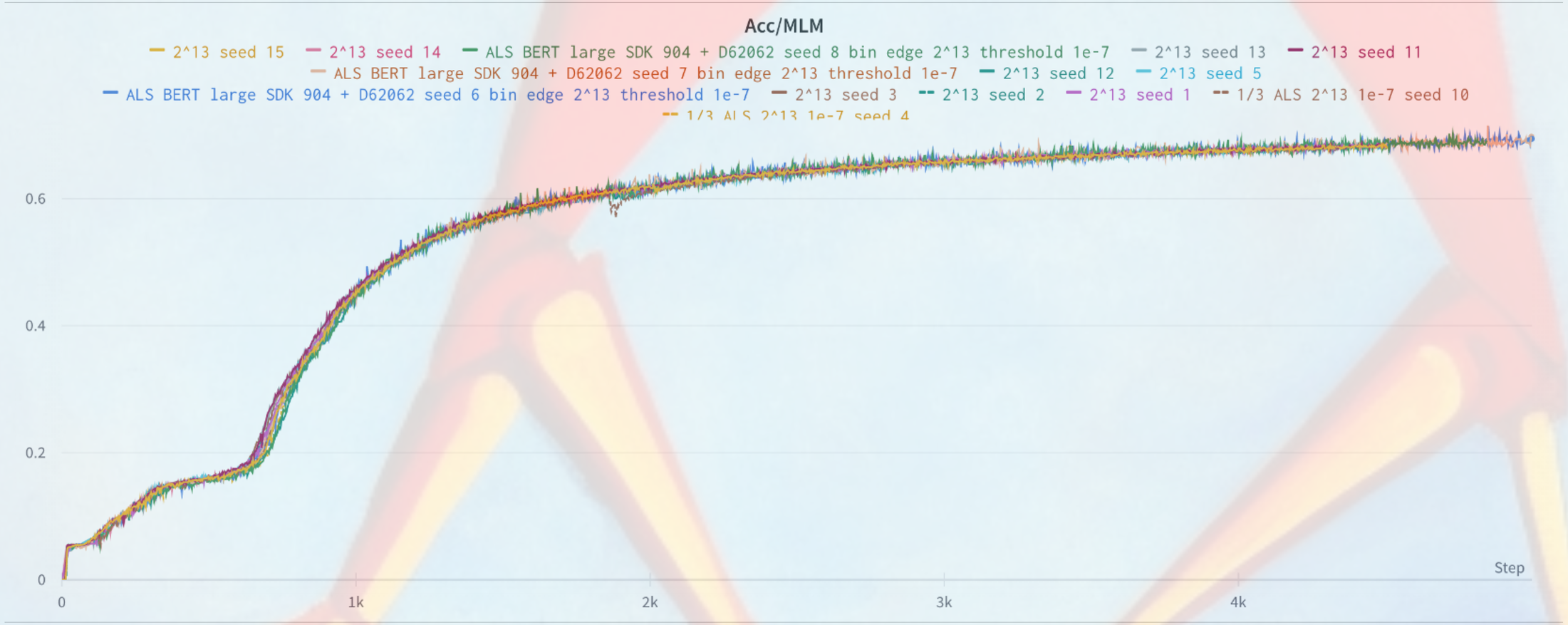




GRAPHCORE

Stable training of BERT Large in reduced precision with ALS

Converging runs with multiple seeds using the ALS method now in production/release



Single line of code in PyTorch models: `opts.Training.setAutomaticLossScaling(True)`

The bad and ugly

Google's school boy error...

npr

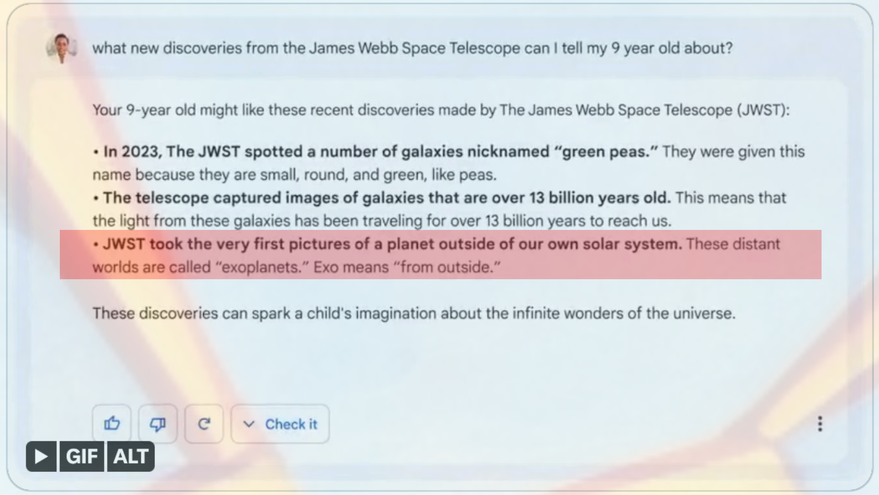
TECHNOLOGY

Google shares drop $100 billion after its new AI chatbot makes a mistake

February 9, 2023 · 10:15 AM ET

Google ✓
@Google

Bard is an experimental conversational AI service, powered by LaMDA. Built using our large language models and drawing on information from the web, it's a launchpad for curiosity and can help simplify complex topics → goo.gle/3HBZQtu



@boelger

Which answer do you prefer?

One that is correct and with sources, or one that confidently makes a mistake?



boelger Today at 2:47 PM
How many solar eclipses can happen in a year?
? 1

NEW

@boelger How many solar eclipses can happen in a year?

brua BOT Today at 2:47 PM
On average, there are two to five solar eclipses each year, with a total eclipse taking place every 18 months or so. Whether you can view that eclipse depends on where you are in the world.
Source:
https://www.britannica.com/science/eclipse/The-frequency-of-solar-and-lunar-eclipses
https://www.space.com/25644-total-solar-eclipses-frequency-explained.html
https://earthsky.org/astronomy-essentials/how-many-solar-or-lunar-eclipses-in-one-calendar-year/
https://www.timeanddate.com/eclipse/how-often-solar-eclipse.html
https://www.nhm.ac.uk/discover/solar-eclipse-guide.html



How many solar eclipses can happen in a year?

In a given year, there can be up to four solar eclipses. However, most years have between two and three solar eclipses. A solar eclipse occurs when the Moon passes between the Earth and the Sun, blocking the light from the Sun and casting a shadow on the Earth. The frequency of solar eclipses varies from year to year because the Moon's orbit around the Earth is not perfectly circular.

@boelger

Tried asking the same question many times?



February 14, 2023

boelger Today at 12:17 AM
How many exoplanets are known to humanity?
❓ 1

@boelger How many exoplanets are known to humanity?

brua BOT Today at 12:18 AM
As of 1 February 2023, there are 5,307 confirmed exoplanets in 3,910 planetary systems, with 853 systems having more than one planet. Most of these were discovered by the Kepler space telescope. There are an additional 2,054 potential exoplanets from Kepler's first mission yet to be confirmed, as well as 978 from its "Second Light" mission and 4,090 from the Transiting Exoplanet Survey Satellite (TESS) mission.
Source:
https://en.wikipedia.org/wiki/Lists_of_exoplanets
https://exoplanets.nasa.gov/faq/6/how-many-exoplanets-are-there/
https://bigthink.com/starts-with-a-bang/first-5000-exoplanets/

boelger Today at 12:27 AM
How many exoplanets are known to humanity?
❓ 1

@boelger How many exoplanets are known to humanity?

brua BOT Today at 12:27 AM
As of 1 February 2023, there are 5,307 confirmed exoplanets in 3,910 planetary systems, with 853 systems having more than one planet.
Source:
https://en.wikipedia.org/wiki/Lists_of_exoplanets
https://exoplanets.nasa.gov/faq/6/how-many-exoplanets-are-there/
https://bigthink.com/starts-with-a-bang/first-5000-exoplanets/

boelger Today at 12:28 AM
How many exoplanets are known to humanity?
❓ 1

@boelger How many exoplanets are known to humanity?

brua BOT Today at 12:29 AM
As of 1 February 2023, there are 5,307 confirmed exoplanets in 3,910 planetary systems, with 853 systems having more than one planet.
Source:
https://en.wikipedia.org/wiki/Lists_of_exoplanets
https://exoplanets.nasa.gov/faq/6/how-many-exoplanets-are-there/
https://bigthink.com/starts-with-a-bang/first-5000-exoplanets/

How many exoplanets are known to humanity?

As of my knowledge cut off in 2021, the number of confirmed exoplanets was 4,134. However, this number is constantly changing as new exoplanets are discovered and confirmed. It is likely that the current number of known exoplanets is higher than that.

How many exoplanets are known to humanity?

As of my training data, which has a cutoff date of 2021, the number of confirmed exoplanets is 4,332. However, this number is constantly changing as new exoplanets are discovered and confirmed by astronomers. It's worth noting that this number only represents the exoplanets that have been confirmed through various methods, and it's estimated that there are many more exoplanets in the universe that have yet to be discovered.



How many exoplanets are known to humanity?

As of my knowledge cut off in 2021, the number of confirmed exoplanets (planets outside our solar system) was around 4,000. However, this number is constantly changing as new discoveries are made and more data becomes available. Astronomers are actively searching for exoplanets using various methods, and it's estimated that there may be tens of billions of exoplanets in our Milky Way galaxy alone.

@boelger

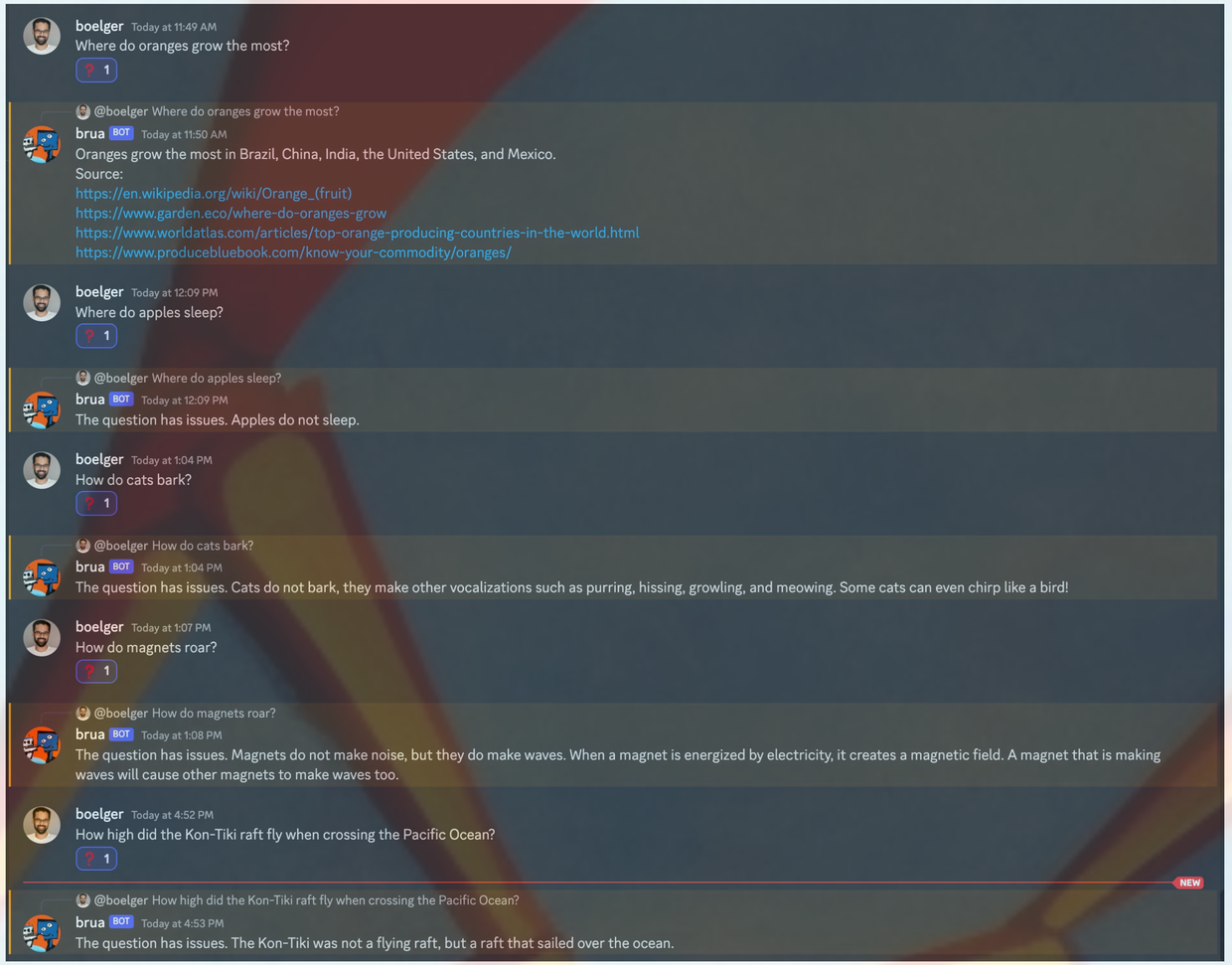
Let’s look at hallucinations in action...

The useful



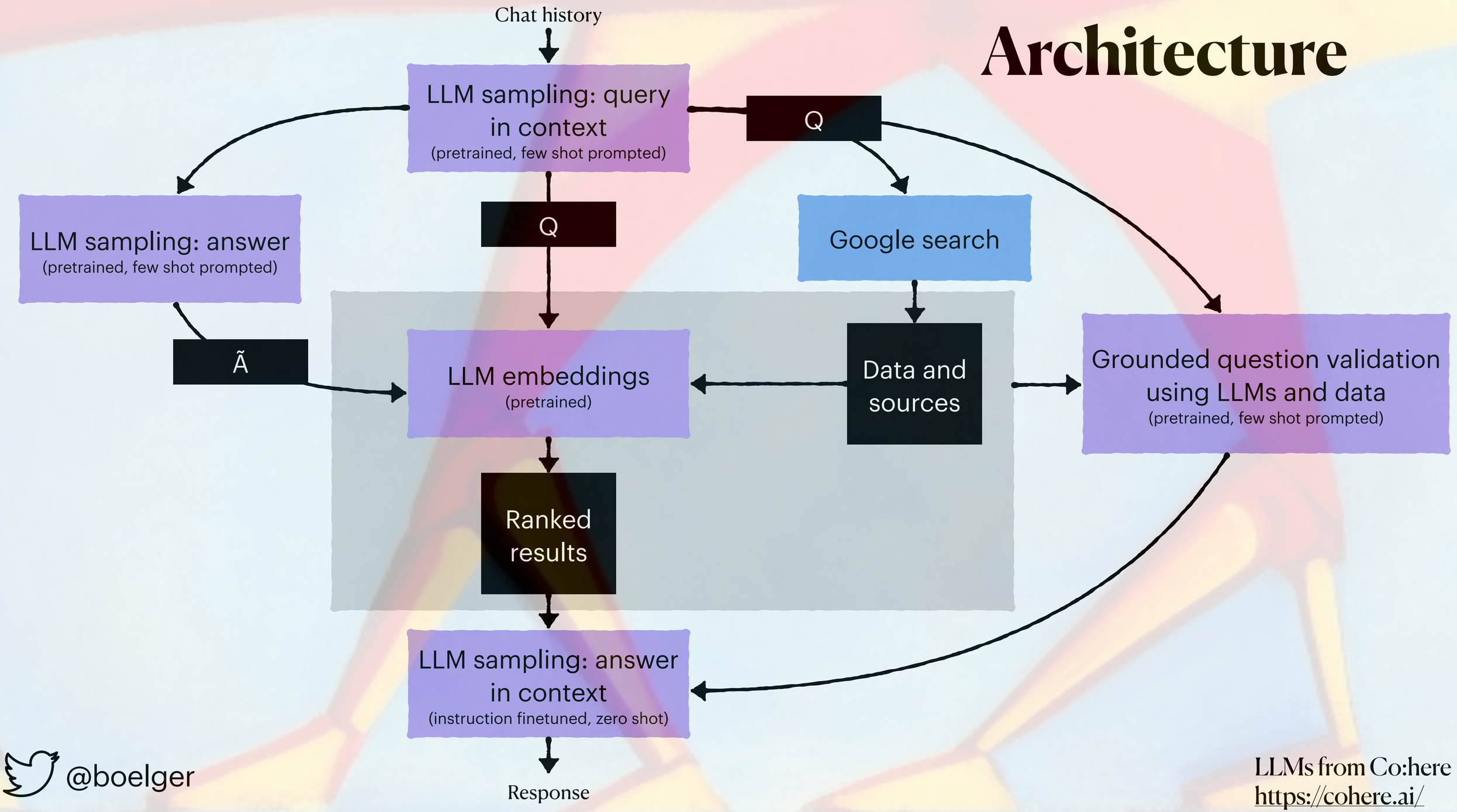
Brua the Critical

Composition of LLMs and an expanding suite of reliability capabilities



@boelger

Architecture



@boelger

LLMs from Co:here
https://cohere.ai/



Keeping AI capabilities on firm ground...

@boelger

Brua  is an expression of brua.io

brua.io is on a mission to transform the way humans interact with AI capabilities. We are here for these capabilities to augment human endeavours reliably and safely, building a deep tech company rooted in Oslo that expands the frontier of AI along the way.

✅ values ✅ talent ✅ tech leadership ✅ vision ✅ ambition

Let's shape the future of AI right here!

@boelger

Thank you!

arjun@brua.io



@boelger