



European Laboratory for Learning and Intelligent Systems

TRAINING AN OPEN-ACCESS 80 BILLION PARAMETERS VISION AND LANGUAGE MODEL

ELLIS Summer school
Large-scale Artificial Intelligence
Modena, Italy 18-22 September 2023

VICTOR SANH
Hugging Face
victor@huggingface.co
@SanhEstPasMoi



UNIMORE
UNIVERSITÀ DEGLI STUDI DI
MODENA E REGGIO EMILIA

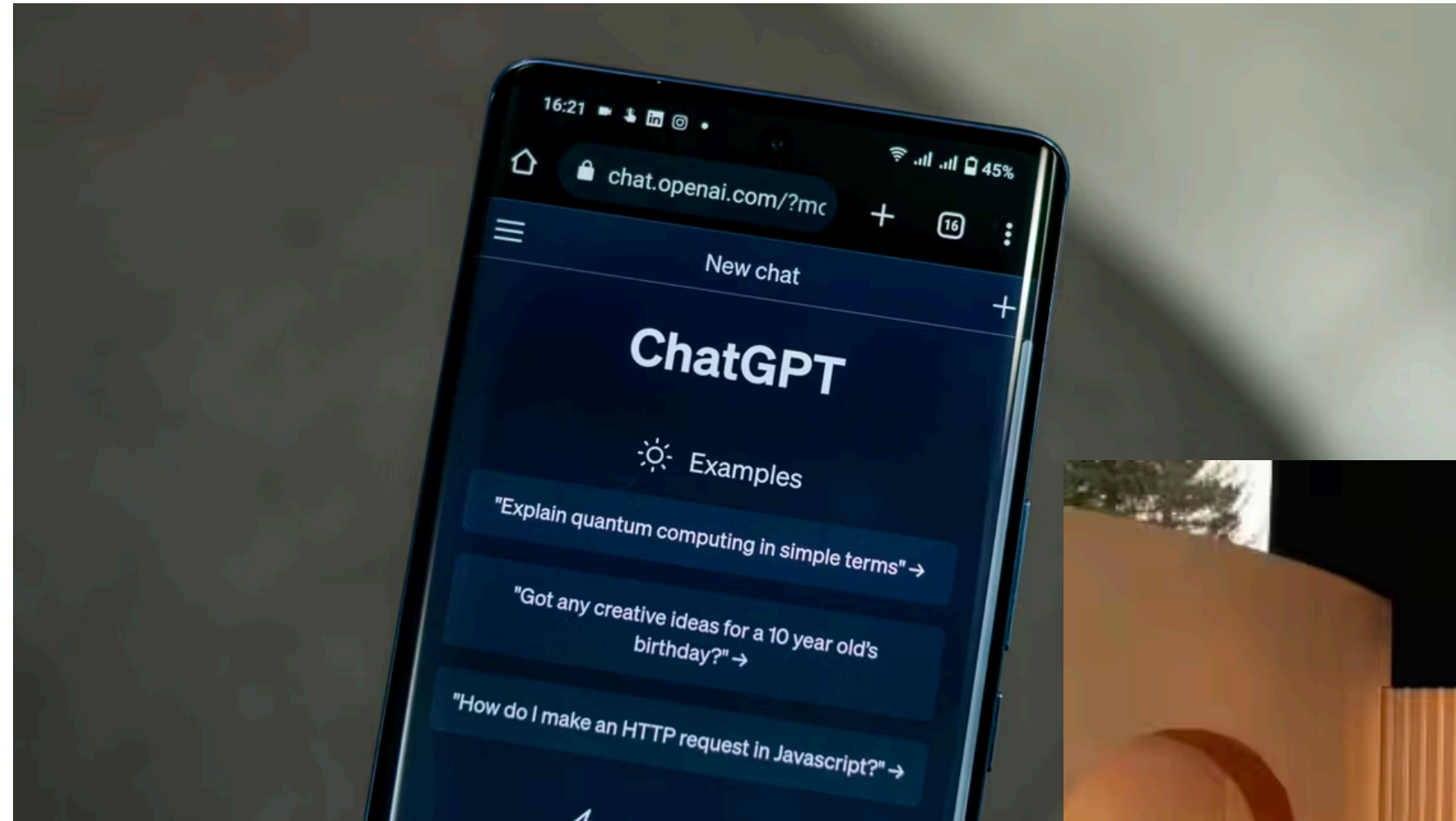
Training an open-access 80 billion parameters vision and language model

Lessons and mistakes

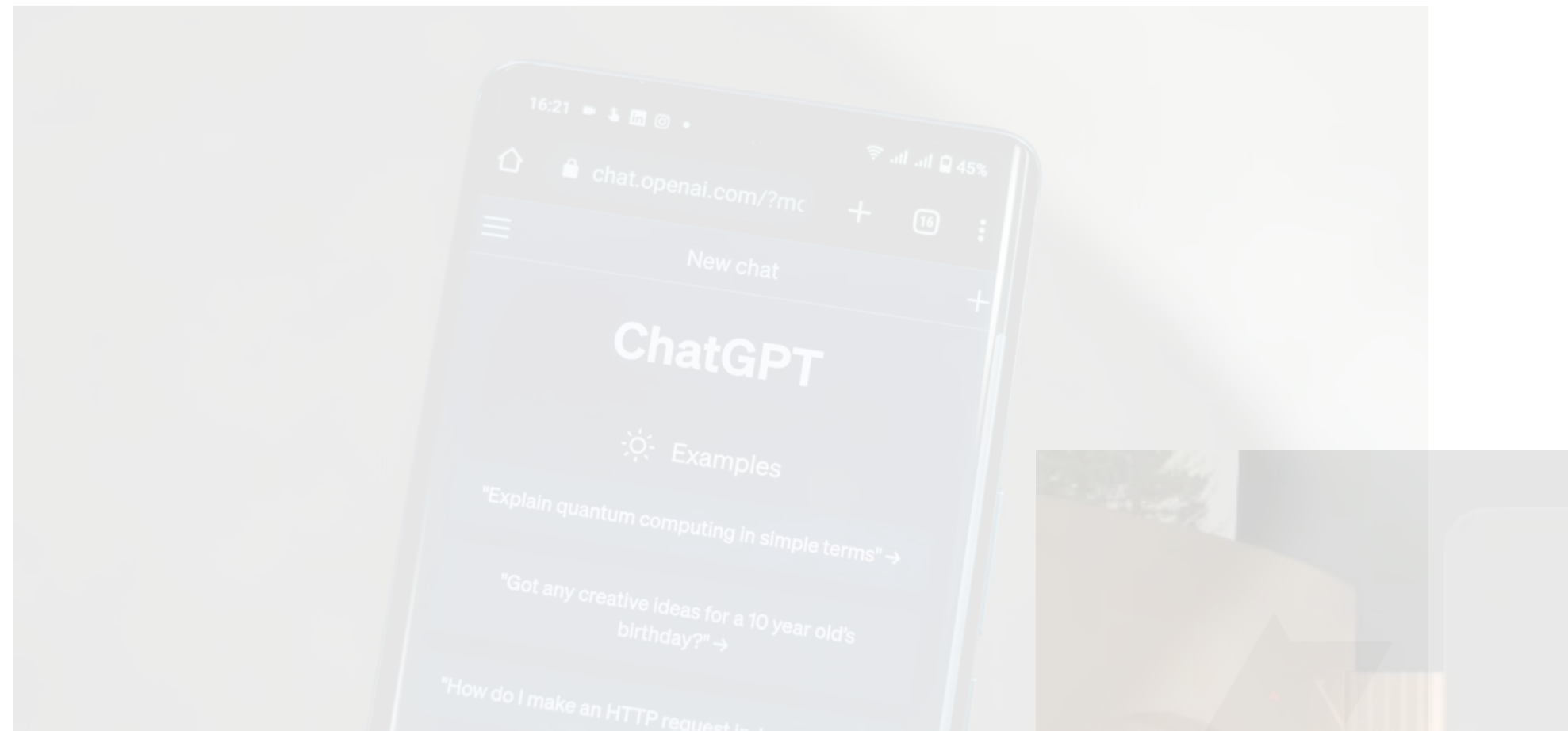
Twitter: @SanhEstPasMoi



LLMs are the talk of the town



LLMs are the talk of the town



Man ends his life after an AI chatbot 'encouraged' him to sacrifice himself to stop climate change

State of the Union: Artificial intelligence start-ups to get easier access to EU supercomputers

14 Sep 2023 | News

In her speech to MEPs, Ursula von der Leyen also called for an equivalent body to the Intergovernmental Panel on Climate Change, with a brief to 'set minimum global standards' for safe and ethical use of AI

By Martin Greenacre



European Commission President Ursula von der Leyen delivering the State of the Union Address on Wednesday. Photo: Christophe Licoppe / European Union



Multimodality: the next frontier of large scale AI systems

> Lots of progress in modalities other than text (in part) powered by Transformers (diffusion models, text2speech, ViT, etc.)

User Can you explain why this is funny. Think about it step-by-step.




GPT-4 The comic is satirizing the difference in approaches to improving model performance between statistical learning and neural networks.

In statistical learning, the character is shown to be concerned with overfitting and suggests a series of complex and technical solutions, such as minimizing structural risk, reworking the loss function, and using a soft margin.


Multimodality: the next frontier of large scale AI systems

IDEFICS




HUGGING FACE


Write a poem about the company in this image



organizing the Woodstock of AI event,



and the fact that they brought those to the event.



In the heart of San Francisco,
A company named Hugging Face,
Brought together the best of AI,
For the Woodstock of AI event,

With attendees from all around,
To share knowledge and ideas,
And to celebrate the power of AI,
With its potential and possibilities,

Multimodality: the next frontier of large scale AI systems

IDEFICS




Who is in this picture? Why do people find it surprising?

The person in the picture is Pope Francis, the head of the Catholic Church. People find it surprising because he is wearing a large, puffy white coat, which is not a typical attire for a pope. The coat is quite eye-catching and unusual, which is why it has attracted attention.



Multimodality: the next frontier of large scale AI systems

IDEFICS



Explain this meme to me.


The meme features a picture of a grumpy-looking cat with the caption "I like French people they eat pain." This is a humorous play on words, as the French word "pain" means "bread" in English, but it sounds similar to the English word "pain" which means "suffering" or "discomfort." The meme suggests that the cat likes French people because they eat bread, which is a food that the cat might enjoy as well. The meme is intended to be lighthearted and amusing, poking fun at the language differences between English and French.

Write another joke about french people and how they treat cats.


A Frenchman walks into a bar with a cat on his shoulder. The bartender says, "Hey, that's a nice cat. What's his name?" The Frenchman replies, "I call him 'Le Petit Fromage' because he's so soft and cheesy." The bartender laughs and says, "That's a great name. What does he eat?" The Frenchman replies, "He eats bread, of course! He's French, after all." The bartender chuckles and says, "Well, I guess that makes sense. Do you mind if I pet him?" The Frenchman replies, "Sure, but be careful. He's a bit of a pain."

Multimodality: the next frontier of large scale AI systems


AI Dad Jokes




Write something funny to go with that image!

 Two dinosaurs on a coffee break discussing their latest TikTok video.

i want another one about these dinosaurs

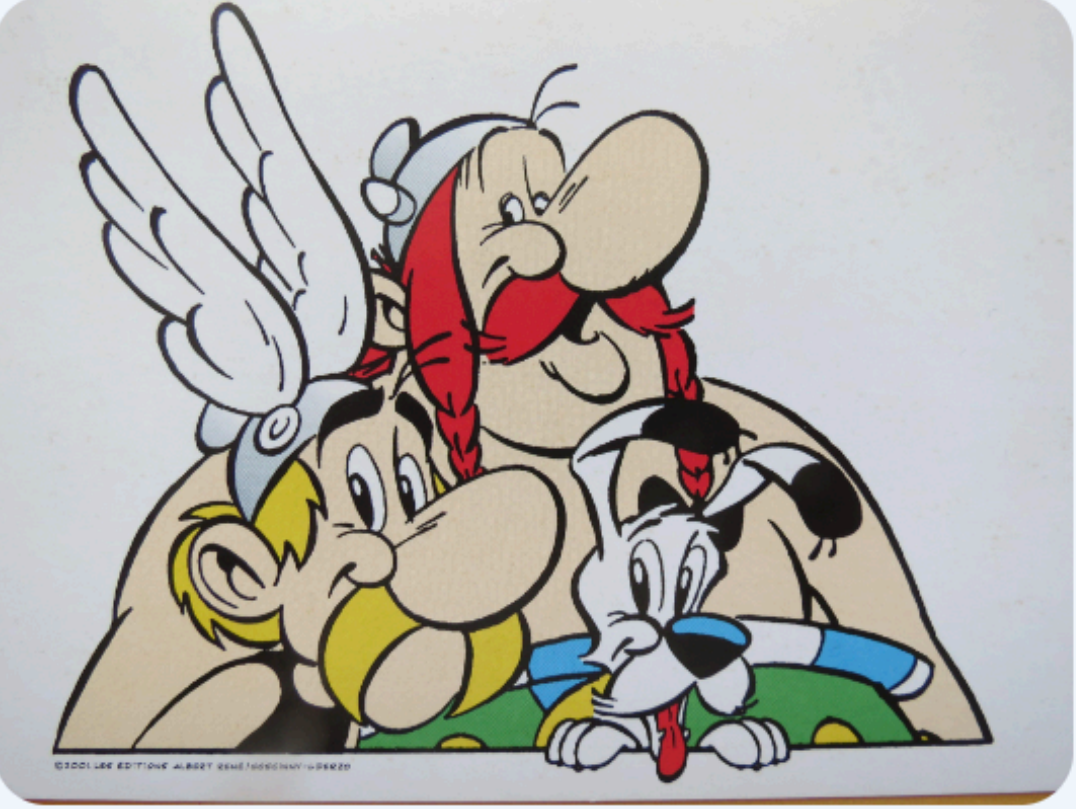
 They both love dancing to 'Dinosaur' by Kesha.

A penguin enters the room


 One of the dinosaurs thinks "Wait, isn't he extinct?"

Multimodality: the next frontier of large scale AI systems

IDEFICS




Can you name the characters in the image and give their French names?

 Yes, the characters in the image are Asterix, Obelix, and Dogmatix. Their French names are Astérix, Obélix, and Idéfix.

Multimodality: the next frontier of large scale AI systems

IDEFICS



Write a complete sales ad for this product.

Introducing the new Pantene Thicker Stronger Hair in just 14 days! Our breakthrough formula with Pro-Vitamin B7 and Keratin Repair Technology helps to strengthen and thicken your hair, leaving it looking and feeling healthier than ever before. With regular use, you'll notice a significant improvement in your hair's texture and volume, making it more manageable and easier to style. So, say goodbye to thin, weak hair and hello to thicker, stronger hair with Pantene!

Rewrite that in italian!

Introducendo il nuovo Pantene Thicker Stronger Hair in soli 14 giorni! La nostra formula rivoluzionaria con Pro-Vitamina B7 e la tecnologia di riparazione del Keratina aiuta a rinforzare e a ingrossare i capelli, lasciandoli apparire e sentire più sani di prima. Con l'uso regolare, noterai un notevole miglioramento nella texture e nel volume dei capelli, rendendoli più maneggevoli e facili da stilare. Dunque, dimentica i capelli sottili e deboli e saluta i capelli più spessi e più forti con Pantene!

Now in german

Multimodality: the next frontier of large scale AI systems

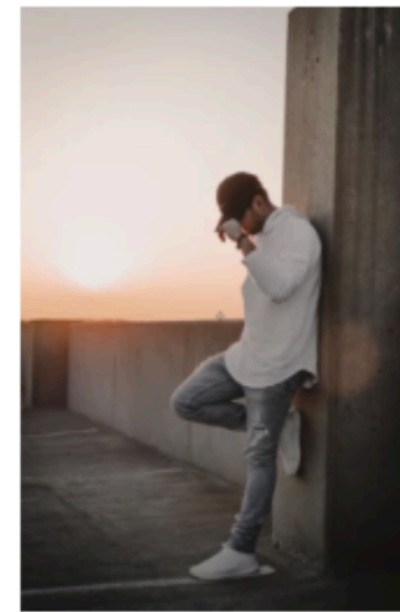


https://huggingface.co/spaces/HuggingFaceM4/idefics_playground

Multimodality: the next frontier of large scale AI systems

> Truly robust end2end multimodal assistants are still a few years ahead

User 🗣️ My friend took a picture of me. Could you edit it to make as if I was standing against the Empire State Building?



AI 🖥️



The system needs to process inputs that span text and images.

The system needs to be able to query external knowledge bases to fetch images of the Empire State building.

User 🗣️ My left foot looks strange. Could you bring the buildings to the front?



User 🗣️ Shadow the visible face of the buildings so that the lightning is coherent with the sun coming from the back.

AI 🖥️



The system needs to process voice inputs.

The system needs to be able to perform a large variety of tasks.

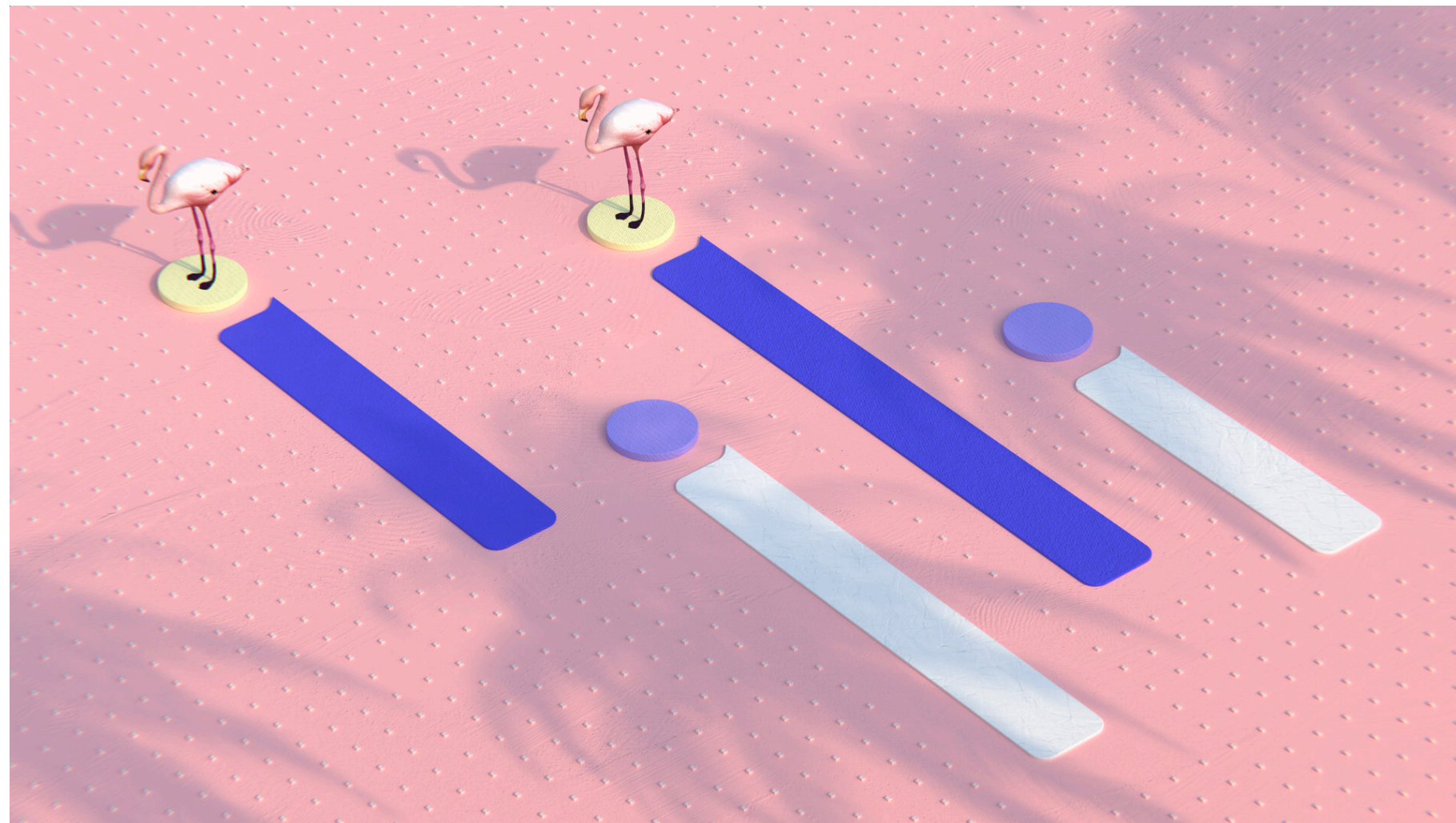
Figure 2: A motivating use case for an adaptable AI system: multimodal iterative editing. The system must master multiple editing skills, process, understand and generate information in several modalities and languages.

The things we don't write in papers: Lessons learned and mistakes

- > You already read papers all week long
- > A presentation that is a bit different than the ones you have seen so far
- > Focussing on the struggles of building a large AI systems
- > Ask questions!!

Program

- > Introduction of Flamingo, a vision and language model
- > IDEFICS, our reproduction of Flamingo
- > Open-questions for future works
- > Multimodal instruction fine-tuning: making these systems useful





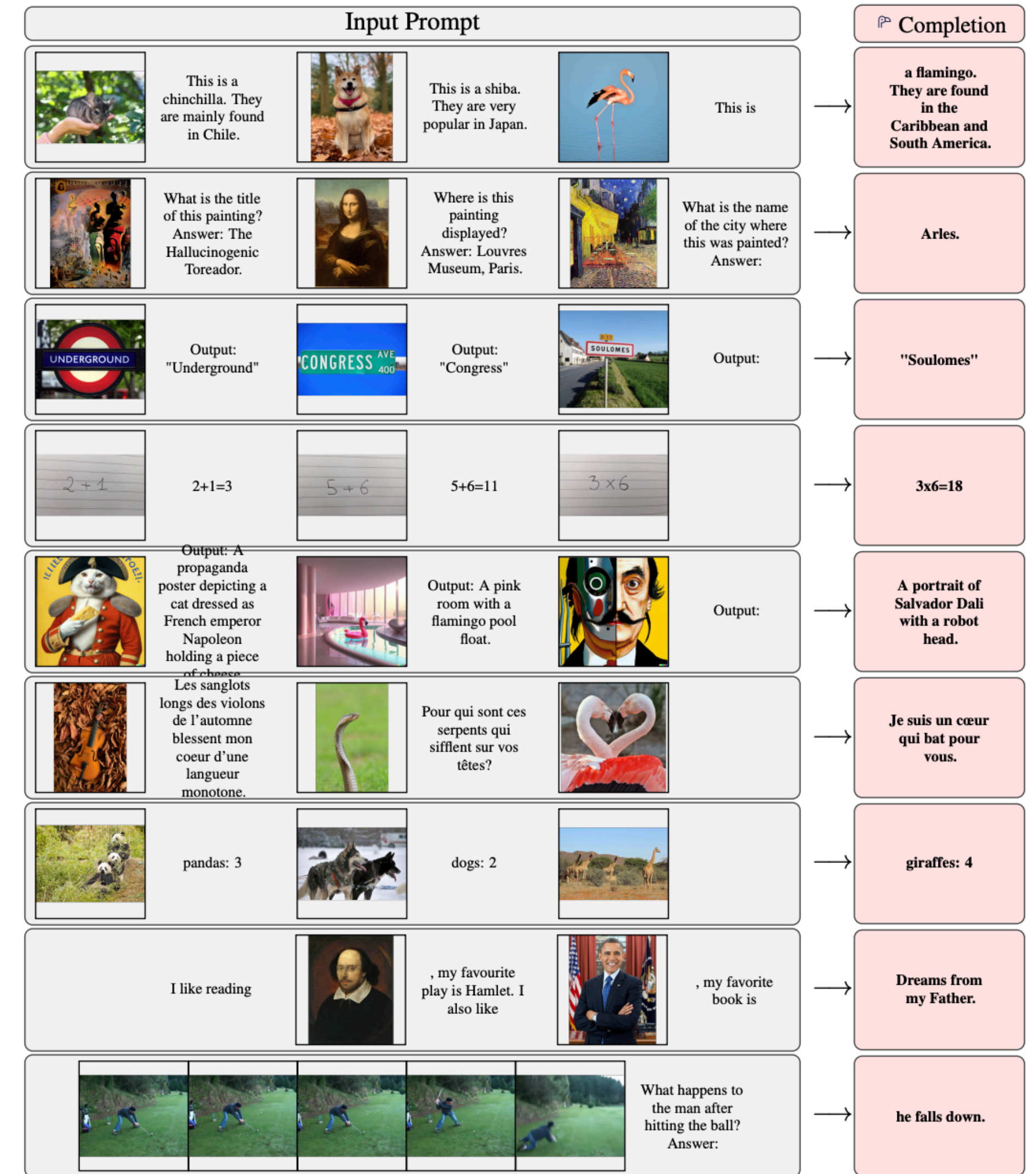
Flamingo: a vision and language model

Flamingo: a vision and language model

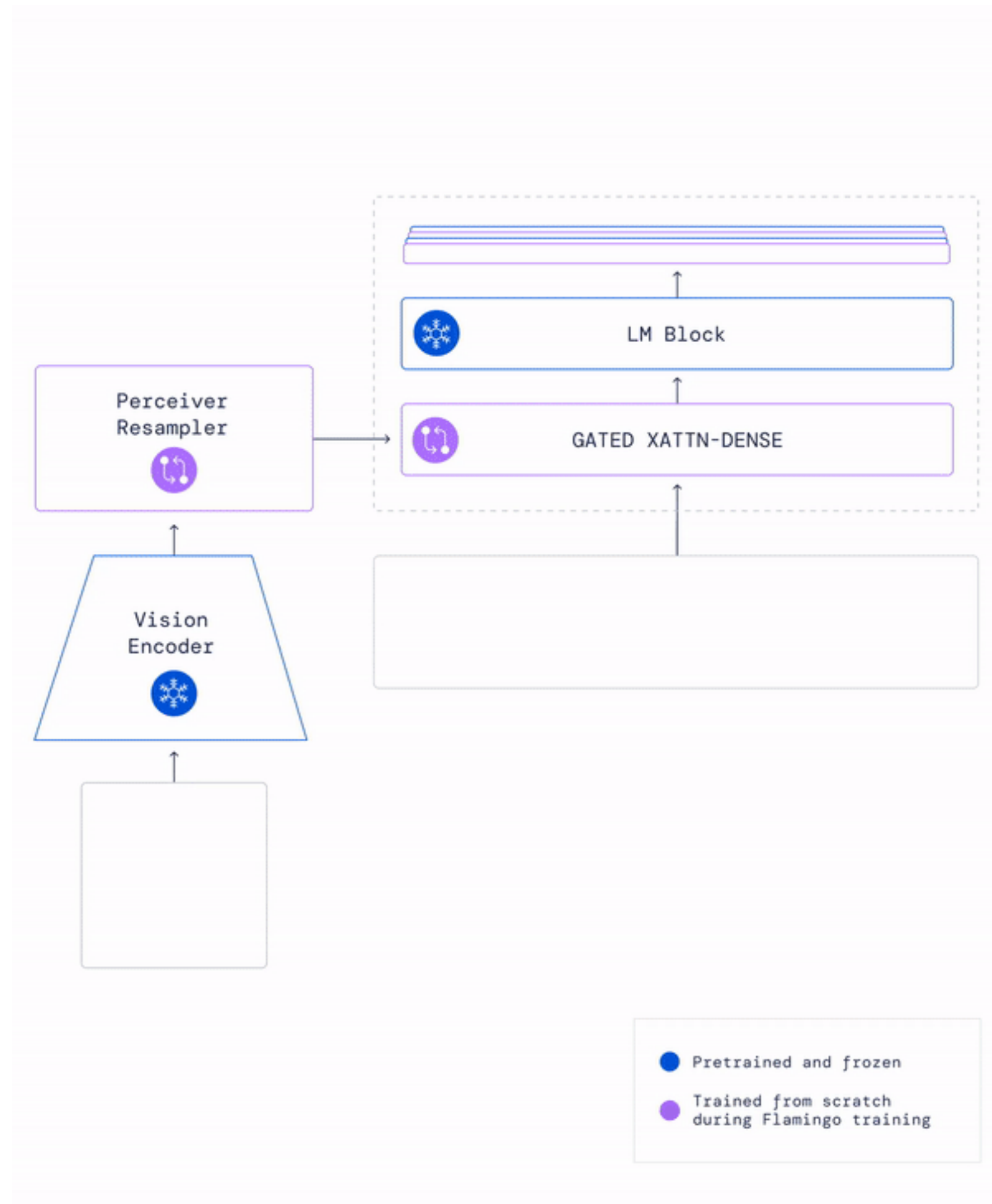
🦩 Flamingo: a Visual Language Model for Few-Shot Learning

Jean-Baptiste Alayrac^{*,‡} **Jeff Donahue^{*}** **Pauline Luc^{*}** **Antoine Miech^{*}**
Iain Barr[†] **Yana Hasson[†]** **Karel Lenc[†]** **Arthur Mensch[†]** **Katie Millican[†]**
Malcolm Reynolds[†] **Roman Ring[†]** **Eliza Rutherford[†]** **Serkan Cabi** **Tengda Han**
Zhitao Gong **Sina Samangooei** **Marianne Monteiro** **Jacob Menick**
Sebastian Borgeaud **Andrew Brock** **Aida Nematzadeh** **Sahand Sharifzadeh**
Mikolaj Binkowski **Ricardo Barreira** **Oriol Vinyals** **Andrew Zisserman**
Karen Simonyan^{*,‡}

<https://arxiv.org/abs/2204.14198>

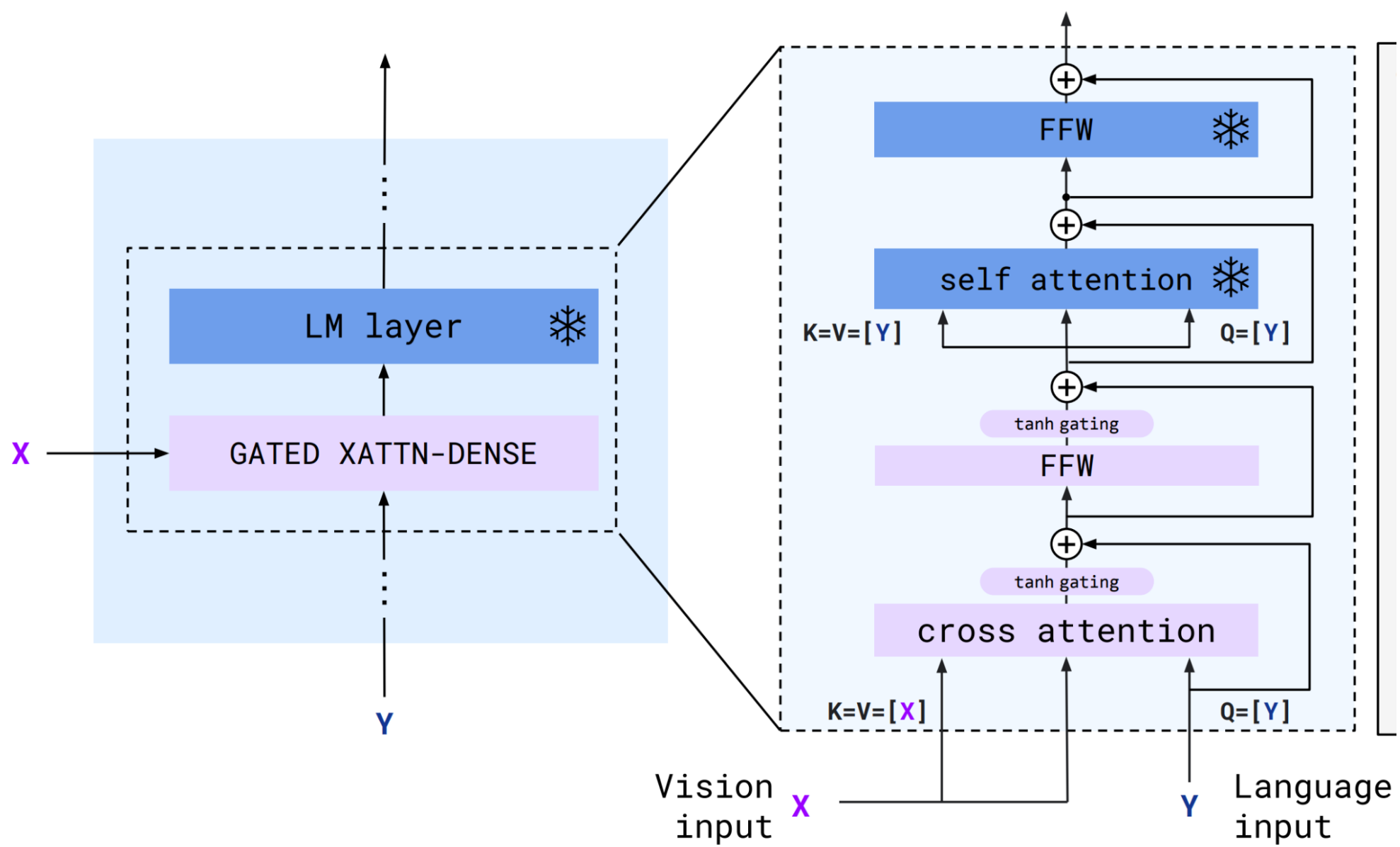


Flamingo: a vision and language model



- > Combines “off-the-shelf” vision & language pretrained models and learn multimodal few-shot in-context learning
- > Frozen LM → Chinchilla
- > Frozen Vision → NFNet-F6

Flamingo: a vision and language model



> Gating \rightarrow Keeps frozen LM “intact” by constraining *amount* of context to use

Flamingo: a vision and language model

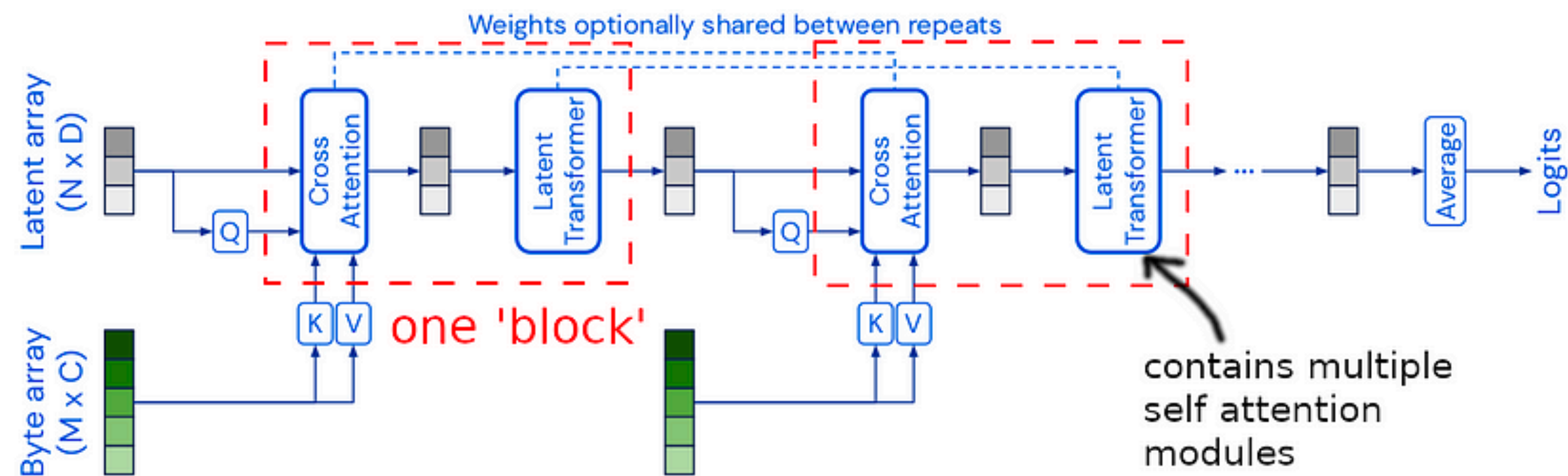
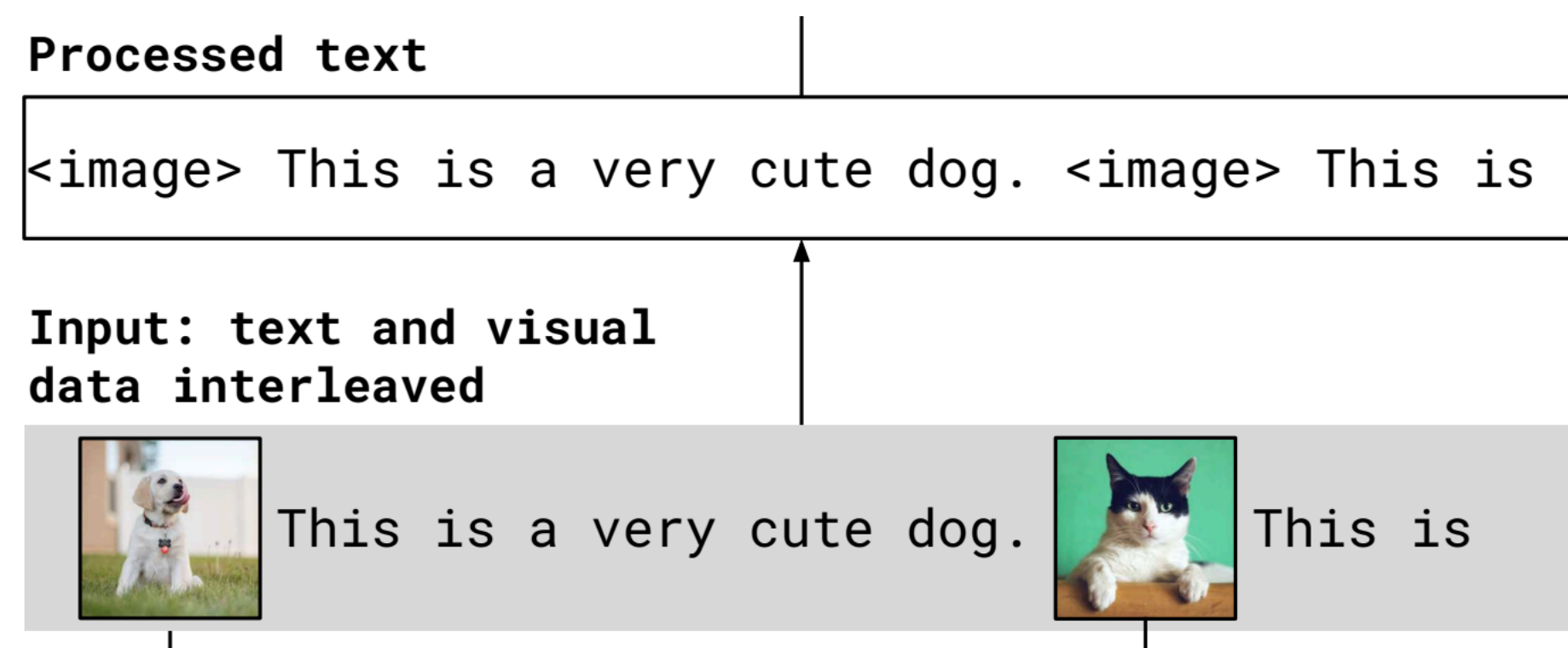


Figure 1. The Perceiver is an architecture based on attentional principles that scales to high-dimensional inputs such as images, videos, audio, point-clouds, and multimodal combinations without making domain-specific assumptions. The Perceiver uses a cross-attention module to project an high-dimensional input byte array to a fixed-dimensional latent bottleneck (the number of input indices M is much larger than the number of latent indices N) before processing it using a deep stack of Transformer-style self-attention blocks in the latent space. The Perceiver iteratively attends to the input byte array by alternating cross-attention and latent self-attention blocks.

<https://arxiv.org/abs/2103.03206>

- > Perceiver → Takes variable number of image or video features from the vision encoder and produces a fixed number of visual outputs.
- > In practice, it reduces the computational complexity (a form of learned pooling)
- > Predefined number of latent input queries

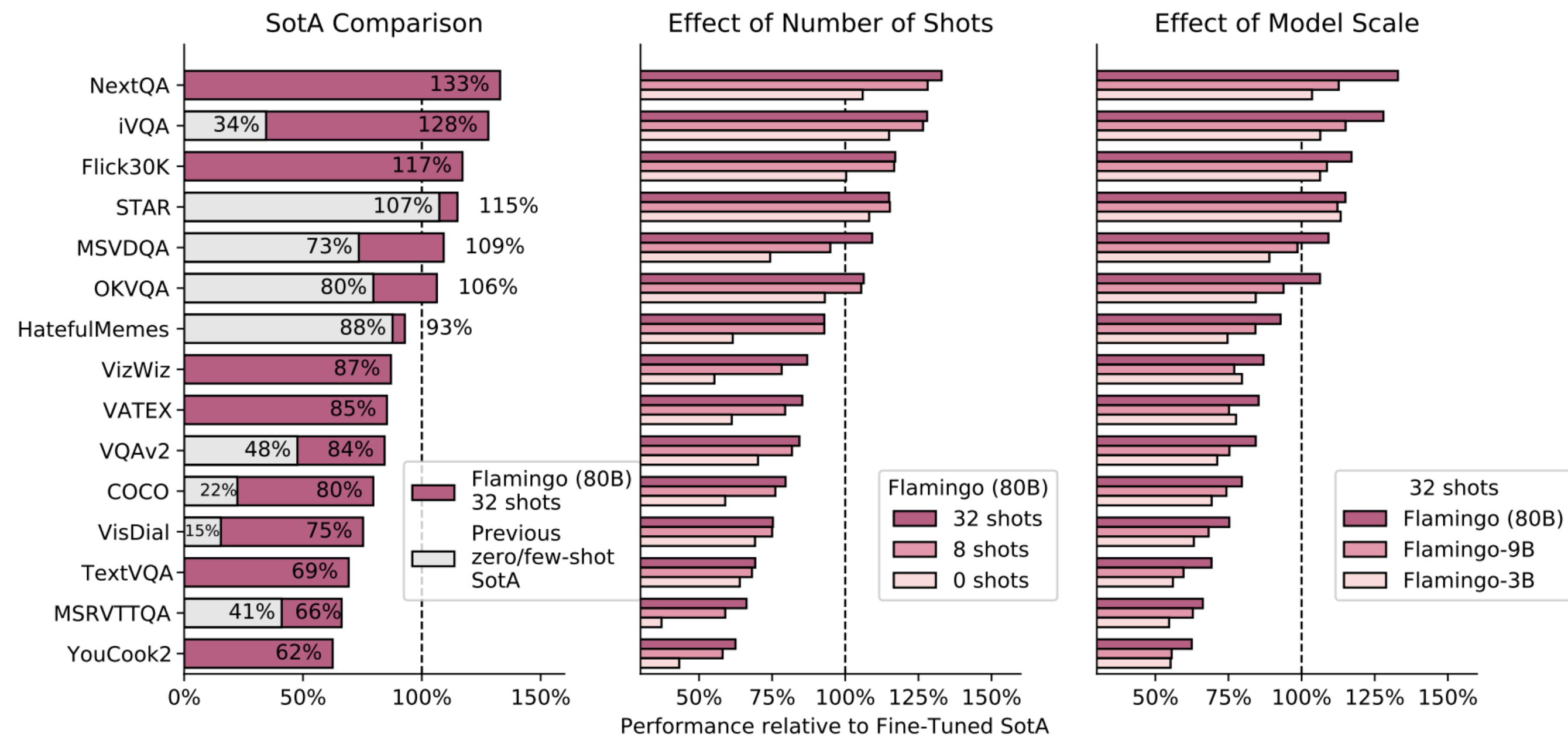
Flamingo: a vision and language model



$$\sum_{m=1}^M \lambda_m \cdot \mathbb{E}_{(x,y) \sim \mathcal{D}_m} \left[- \sum_{\ell=1}^L \log p(y_\ell | y_{<\ell}, x_{\leq \ell}) \right],$$

- > The model is fed arbitrary sequences of text and images and generated text
- > The training objective is the standard next token prediction (same as pure text LM)
- > The model does NOT generate images

Flamingo: a vision and language model



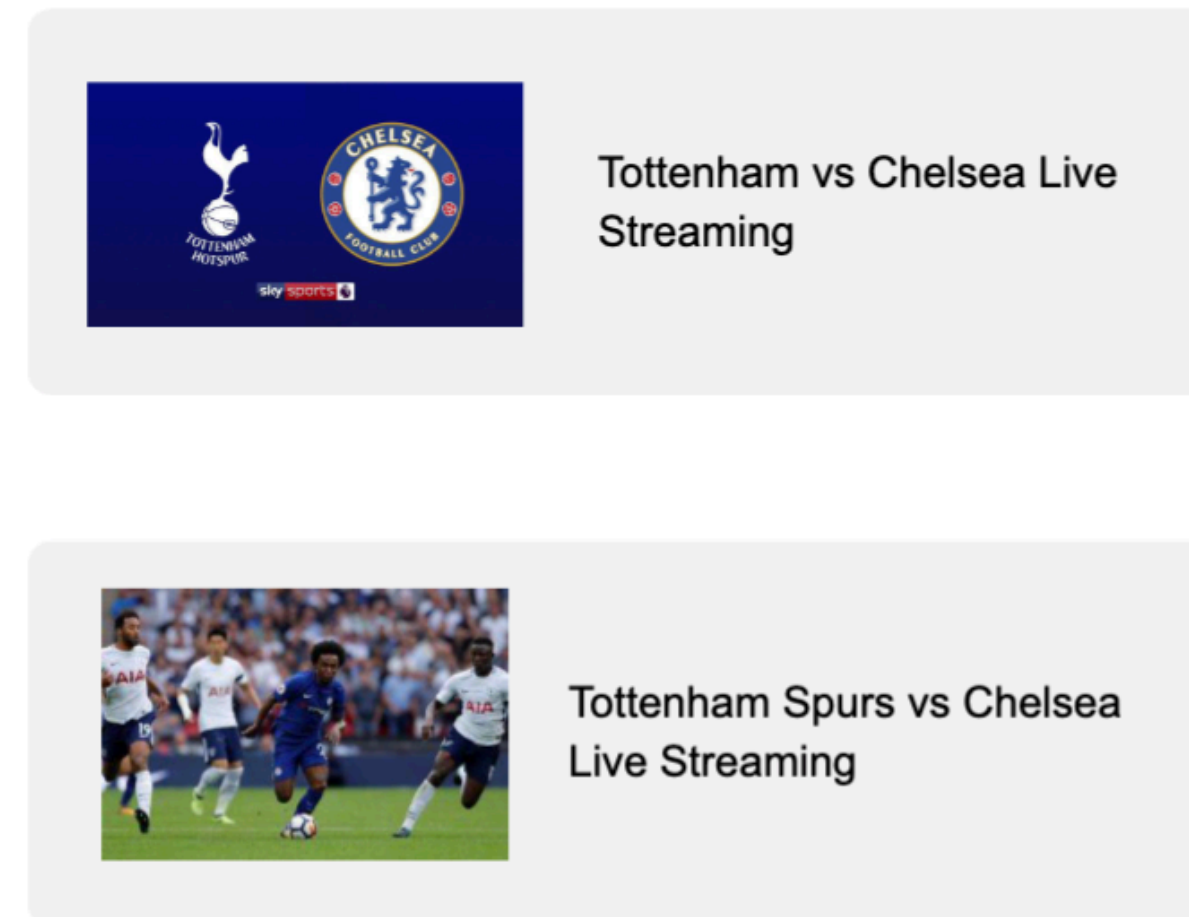
> Emergence of multimodal in-context few-shots abilities

Figure 2 | **Overview of the results of the Flamingo models.** *Left:* Our largest model, dubbed *Flamingo*, outperforms state-of-the-art fine-tuned models on six out of the 16 tasks we consider despite not using any fine-tuning at all. For all 16 tasks where published few-shot results are available, *Flamingo* outperforms them by a large margin and sets the new few-shot state of the art. *Center:* *Flamingo* performance improves with the number of shots. *Right:* The performance of the Flamingo models increases with the model scale. *Note:* We omit RareAct, our 16th benchmark, as it is a zero-shot benchmark with no available fine-tuning results.

Flamingo: a vision and language model

- > **M3W**: 43 million interleaved web pages
- > **ALIGN**: 1.8 billion image-text pairs (alt-text)
- > **LTIP**: 312 million image-text pairs (little details)
- > **VTP**: 27 million of short video-text pairs

Image-Text Pairs



Multimodal Document

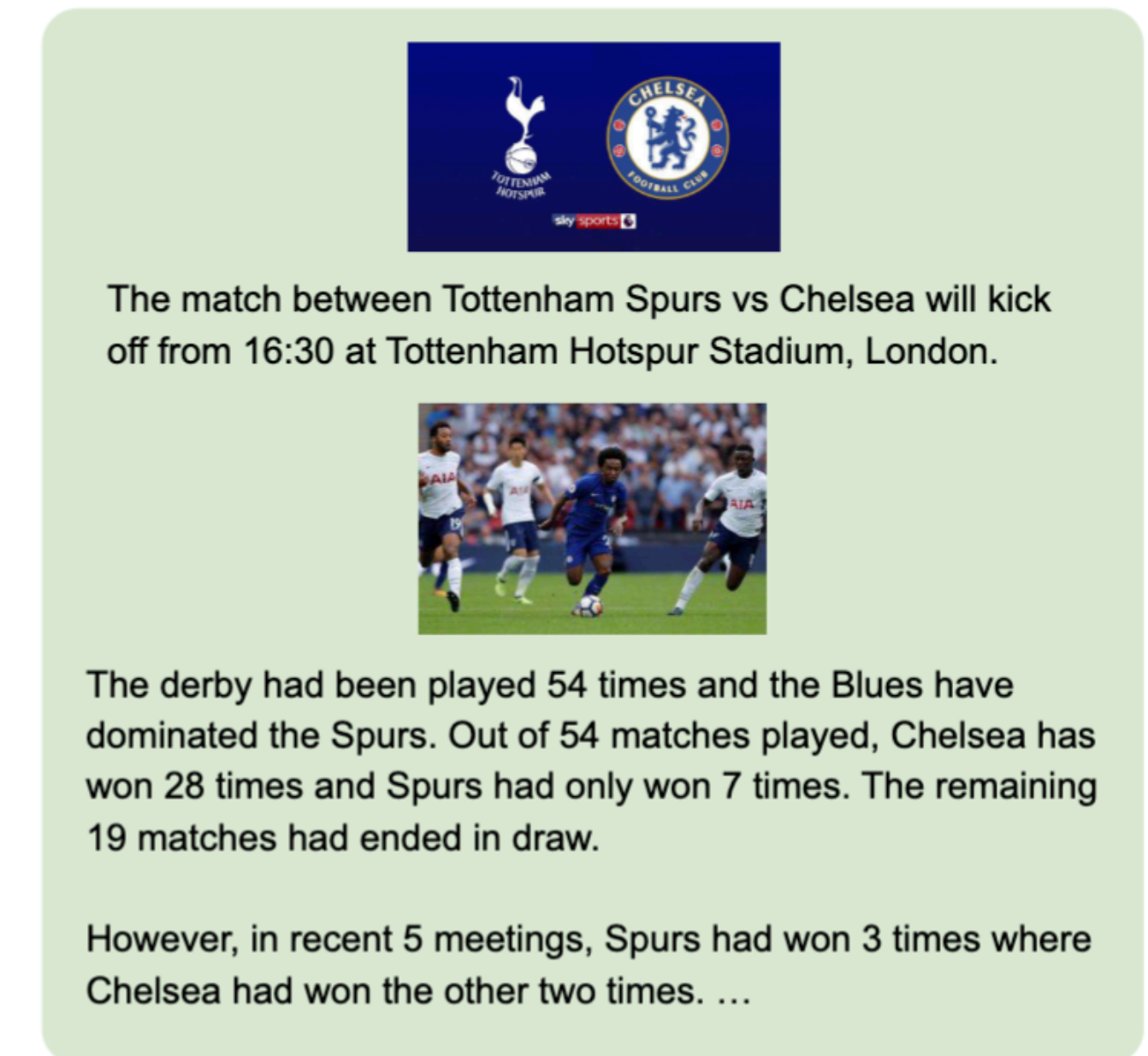


Figure 1: A comparison of extraction from the same web document. For image-text pairs, the alt-text of images is often short or non-grammatical. For OBELICS, the extracted multimodal web document interleaves long-form text with the images on the page.

Flamingo: a vision and language model

Ablated setting	Flamingo 3B value	Changed value	Param. count ↓	Step time ↓	COCO CIDEr↑	OKVQA top1↑	VQAv2 top1↑	ImageNet top1↑	MSVDQA top1↑	VATEX CIDEr↑	Kinetics top1-top5↑	Overall score↑
Flamingo 3B model (short training)			3.2B	1.74s	86.5	42.1	55.8	59.9	36.3	53.4	49.4	68.4
(i) Training data	All data	M3W	3.2B	0.68s	58.0	37.2	48.6	35.7	29.5	33.6	34.0	50.7
		w/o VTP	3.2B	1.42s	84.2	43.0	53.9	59.6	34.5	46.0	45.8	65.4
		w/o LTIP/ALIGN	3.2B	0.95s	66.3	39.2	51.6	41.4	32.0	41.6	38.2	56.5
		w/o M3W	3.2B	1.02s	54.1	36.5	52.7	24.9	31.4	23.5	28.3	46.9



IDEFICS: Open-reproduction of Flamingo

OBELICS: A web-scale interleaved dataset

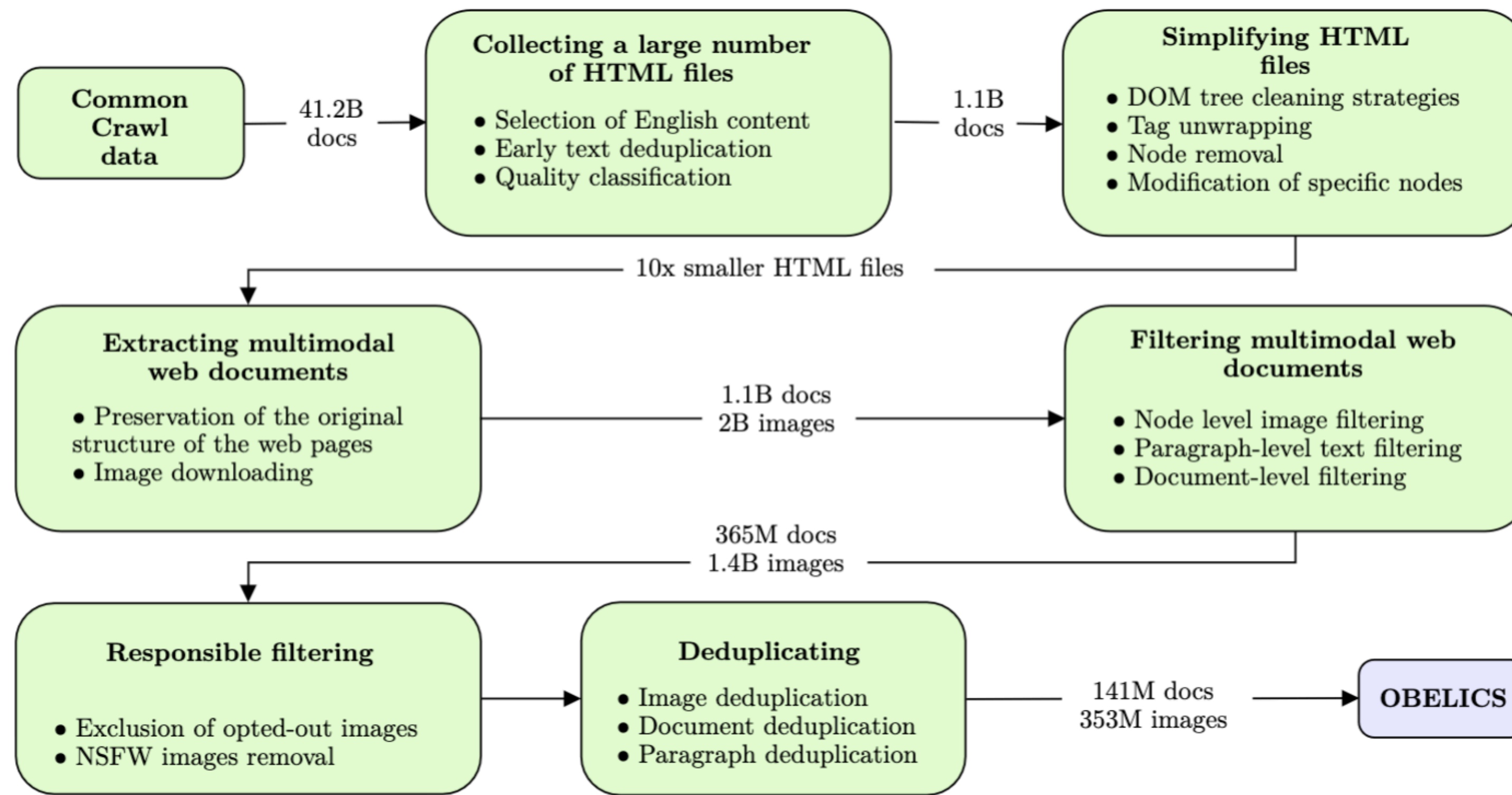


Figure 2: Overview of the steps involved in creating OBELICS.

<https://arxiv.org/abs/2306.16527>

OBELICS: Open Bimodal Examples from Large filtered Commoncrawl Snapshots

> Interleaved image and text

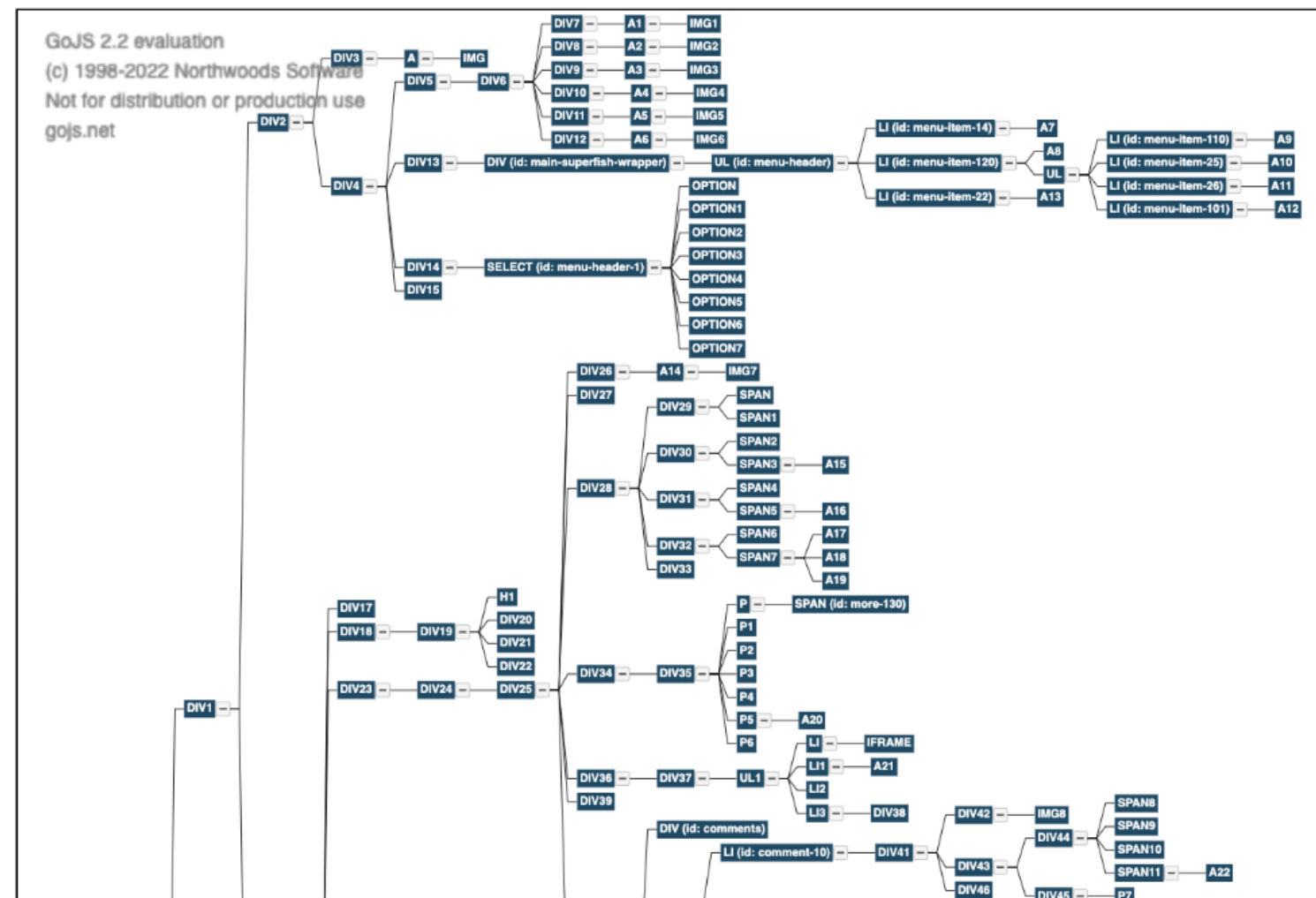
> 141 million web documents

> 353 million associated images

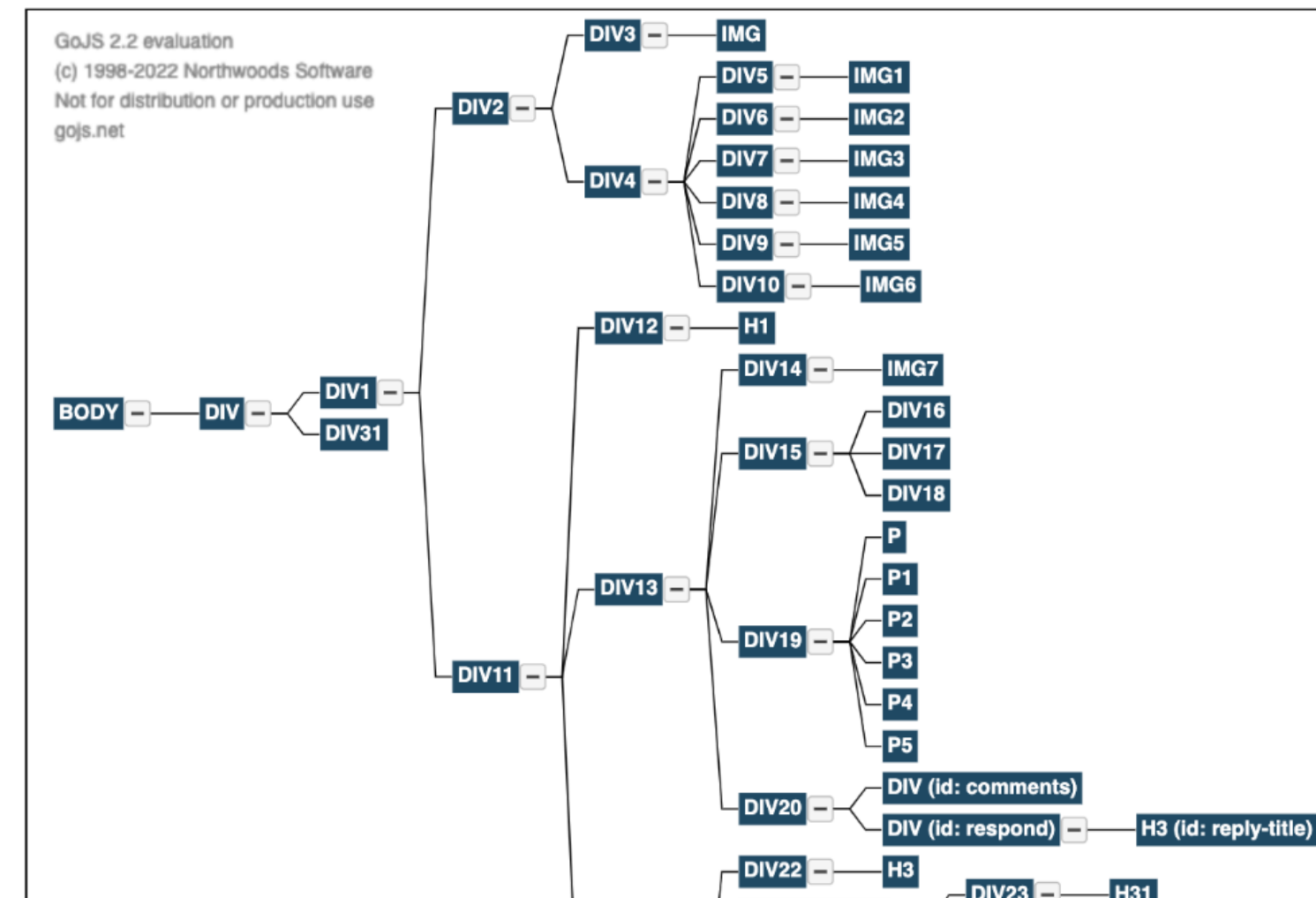
> 115 billion text tokens

OBELICS: A web-scale interleaved dataset

Raw DOM tree



Simplified DOM tree



- > Simplifying HTML files
- > See paper for more details and analysis

OBELICS: A web-scale interleaved dataset

Livestock Guarding Dogs

WELGEVONDEN PROGRAM

Welgevonden Game Reserve has researched leopards in the reserve as part of a greater study to attempt to understand leopard ecology and conservation issues in the Waterberg and elsewhere.



Leopard research is continuing on Welgevonden in association with PANTHERA, Limpopo Department of Environmental Affairs and Tourism (LEDET), University of Kwa-Zulu Natal. Please [click here](#) for more details on the "Limpopo Leopard Project"

The Waterberg Biosphere Reserve is situated within the Waterberg Mountains in the Limpopo Province is well known for its larger conservation areas which include the Marakele National Park as well as privately owned nature reserves, such as Welgevonden Private Game Reserve.

Leopards, and other predators, have existed in these areas naturally despite persecution by livestock farmers and being hunted by trophy hunters.

In addition to Welgevonden, surveys are being conducted in Makalali Game Reserve, Timbavati Game Reserve, Venetia-Limpopo, Atherstone and Wonderkop Nature Reserves, and the western Soutpansberg mountains on non-protected land.

Welgevonden Game Reserve has researched leopards in the reserve as part of a greater study to attempt to understand leopard ecology and conservation issues in the Waterberg and elsewhere.



The Waterberg Biosphere Reserve is situated within the Waterberg Mountains in the Limpopo Province is well known for its larger conservation areas which include the Marakele National Park as well as privately owned nature reserves, such as Welgevonden Private Game Reserve.

Leopards, and other predators, have existed in these areas naturally despite persecution by livestock farmers and being hunted by trophy hunters.

In addition to Welgevonden, surveys are being conducted in Makalali Game Reserve, Timbavati Game Reserve, Venetia-Limpopo, Atherstone and Wonderkop Nature Reserves, and the western Soutpansberg mountains on non-protected land.

This initial phase of this research is a 3 year project (starting in 2013), and Welgevonden was asked to participate in this research due to our history of leopard research.

A conflict of interests therefore exists where leopards, and other predators, are present in the same area as livestock.

OBELICS: A web-scale interleaved dataset

Right now, in Costa Rica, the classic dry season has been evasive. As the sky clouds over just as it did during June, and the rains begin to fall, it almost feels like the whole usual dry season thing has been waived. Cold fronts continue to arrive and subsequently douse the country with Atlantic showers while a "Nina" effect over in the Pacific has only added to the wet situation. Despite the umbrella test, there are good things associated with this. High biodiversity is correlated with high rainfall and that makes for more birds. It's one of the main reasons why so many species occur in Costa Rica.

It can be a challenge to find them under varying degrees of precipitation but what's a birder gonna do? It's part of the local birding scene and when the clouds take a lunch break, the birds suddenly come out to play. Get enough of those breaks and you can get into some stellar birding, especially when high rainfall earlier in the year encouraged the trees and bushes to grow lots of bird friendly fruit. Seriously, it's a smorgasbord out there right now, the tanagers, manakins, thrushes, trogons, and toucans are going to feed whether it rains or not.



When the sun eventually does come out, there seem to be certain birds that take advantage of the sudden bloom of warmth and UV rays. Yesterday morning at El Tapir, a client and myself bore witness to what can happen when the rain finally comes to a stop and the sun, unhindered by clouds, punctuates the sky. At first, there was little activity, as if the birds were still numbed by the constant falling of water, still in denial that the rain had stopped. A few wrens and some other birds vocalized, a pair of Mealy Parrots fluttered overhead but pretty quiet otherwise. However, while the birds of the forest slowly came back to life, the Rufous-tailed Hummingbirds were racing around the garden. Judging by their frantic behavior (even for hummingbirds), it seemed like they hadn't eaten quite enough in days. Or maybe they just didn't get their fill of nectar? Whatever the case, they were drinking from the Verbena flowers as if they were participants in some avian Bacchus festivities. Unfortunately, they didn't invite any other hummingbirds to the party and took great efforts to bounce any potentially crashing woodnymph, Snowcap, or Violet-headed.



Dressed for the party, still denied entrance. Name's not down, not coming in.

It took a while but the Rufous-tailed seemed to eventually get their fill (or became too inebriated) and as the sun took over the garden space, a couple other hummingbird species braved the post party scene. One of the most cooperative was a male Black-crested Coquette.



As is typical with coquettes, the male chose to perch on a bare twig for extended periods of time before carefully flying down to drink from the Verbena. Much to our satisfaction, this particular exquisite beauty preferred to feed on a bush right in front of us.

The Marvel Cinematic Universe has created some magnificent things over the last decade and a half. This cinematic universe has brought them back from the cusp of bankruptcy and into times of abundance once again. The success of the MCU has now allowed Marvel Studios to bring out the obscure characters from comic pages onto the silver screen. Who would have thought that Kit Harrington would be playing Dane Whitman in the MCU? It is relevant because Dane Whitman will become Black Knight, the greatest swordsman on the planet who fights alongside Avengers.



Who is this Black Knight? Why do we care? And why are we talking about this after a movie about cosmic beings like the Eternals and the Celestials? Does a sword not seem moot in front of infinite cosmic energy? Not when it is this sword. You see, in the after-credits scene of Eternals, Dane Whitman aka the love interest of Sersi unveils a sword. This sword seems to whisper to him and looks like the cursed Ebony Blade from the comics. Dane Whitman in the comics wields this blade and calls himself the Black Knight, a superhero who assists the Avengers in various battles.

But there is a catch. The Ebony Blade was supposed to be welded by the pure of heart as explained by Merlin who created the sword. But the secret of the sword is that it can only be wielded by those who are impure of heart. The blade was actually designed by Merline for Sir Percy (ancestor of Dane Whitman) to make him the greatest swordsman at the time. But the catch is that the blade seeks out evil inside you and amplifies it until there is nothing but a berserker left.

This seems to be true in the MCU too. The Ebony Blade blesses its user with incredible power, but it also comes at an incredible cost. This sword also prolongs its user's life as much as it can. The last Black Knight before Dane Whitman was Nathan Garrett, his uncle who is mentioned in the movie several times. This Black Knight was a villain who was defeated by the Avengers in the comics. But here, he is nowhere to be seen. There is a reason for this and the reason is most likely that Nathan Garrett will work better as a villain against Dane Whitman than the Avengers of the MCU.




This Ebony Blade is a malicious piece of weaponry. It was created by Merline so that Sir Percy may sully his honor in battle but it also gave him immense power in the series. There is a possibility that we will see a similar story play out with Kit Harrington's character in the MCU. Moreover, there is another question that we must address. Who does the voice at the end of the second after-credits scene belong to? It has been confirmed by Chloe Zhao that it is Mahershala Ali's Blade who has come to recruit Dane.

OBELICS: A web-scale interleaved dataset

NOMIC

atlas_id_ >
9c05cd07-a0f9-4c82-aebf-c769e8ac30ff
document >

Five Critical and Important Things For New Managers to Remember



Five Critical and Important Things For New Managers to Remember

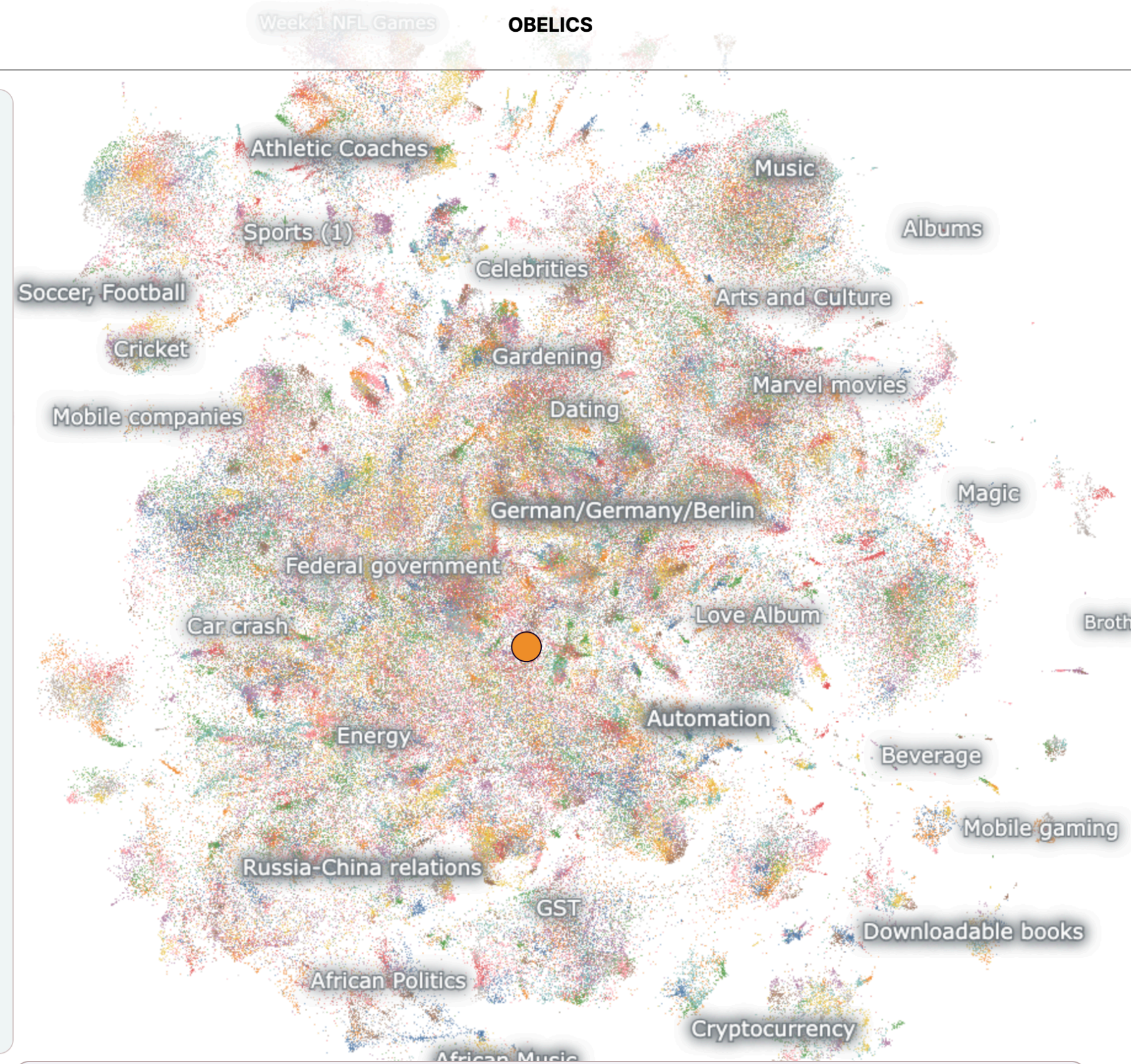
So, you've been promoted to a manager position and have people reporting directly to you. Congratulations! But with great power comes great responsibility.

No matter how successful you've been previously in your professional life, managing people for the first time can be a daunting task.

Here are 10 things to keep in mind to start you off on the right foot.

1. This is not an elevated version of your old job

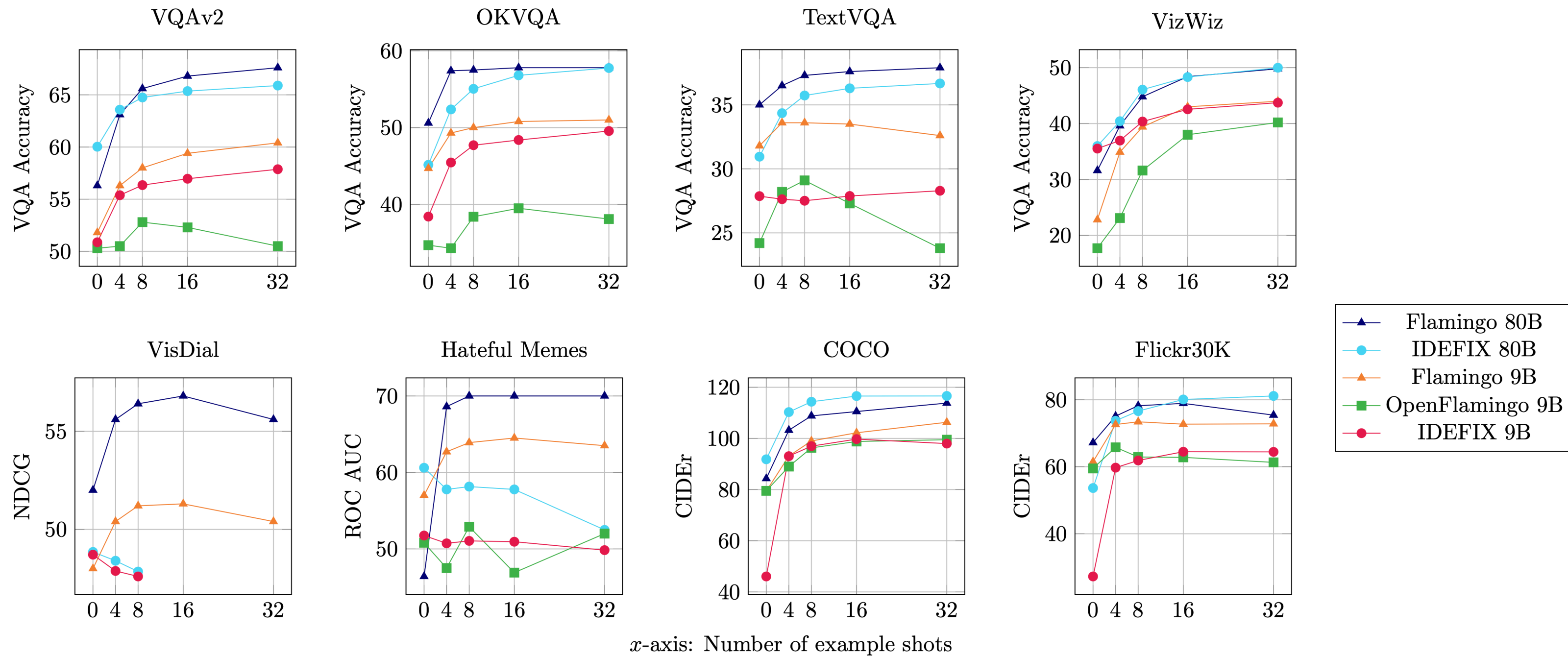
Remember the Peter Principle – idea that you'll



Content Warning x
This map may contain sensitive, inappropriate, or offensive content. Please proceed with caution and exercise discretion while sharing or viewing.

<https://atlas.nomic.ai/map/f2fba2aa-3647-4f49-a0f3-9347daeee499/ee4a84bd-f125-4bcc-a683-1b4e231cb10f>

IDEFICS: Open-reproduction of Flamingo



> IDEFICS 80B is on par with Flamingo 80B
> IDEFICS 9B is not as good as Flamingo 9B

IDEFICS: Open-reproduction of Flamingo

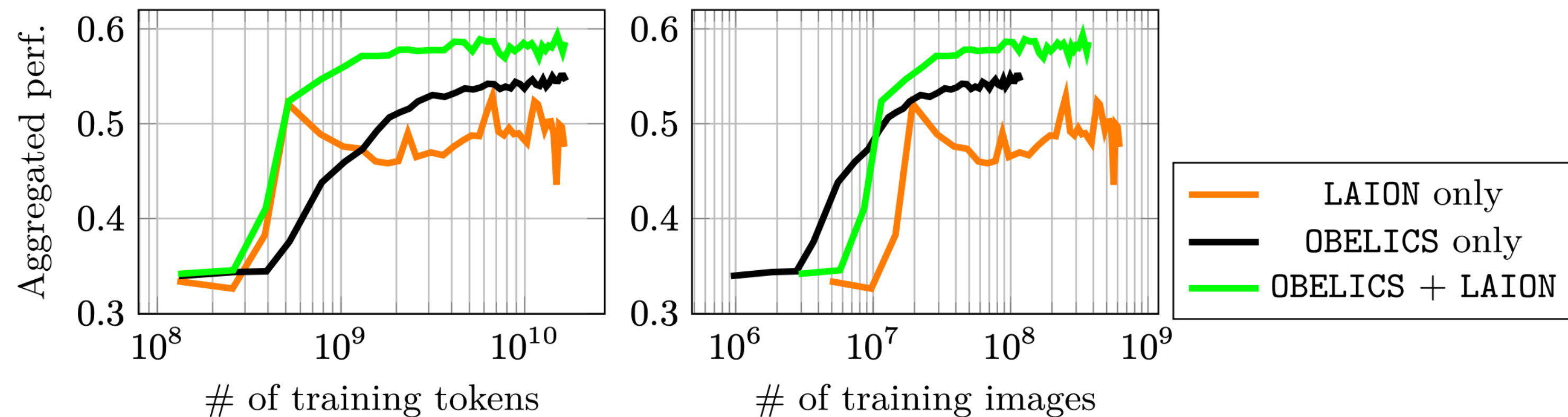



Figure 6: Aggregated performance through the training using LAION only, OBELICS only and a mixture of both. The training sequences from multimodal documents and the packed sequences obtained from image-text pairs have different numbers of images but the same number of tokens. Thus, we plot the performance over two log x-axes. The initial uptick of the model trained on image-text pairs is attributed to the fact the performance on VQA tasks starts by increasing and then slowly degrades.

- > Training on multimodal web docs allows reaching the same performance using 10x fewer images than training on image-text pairs and yields better performance
- > Striking on VQA
- > Worse on OCR, Image captioning, Classification


IDEFICS: Open-reproduction of Flamingo

IDEFICS




Describe this image.

The image shows a robot with a square head and a red light on its chest. It has a yellow arm and a blue arm, and it is holding a small box. The robot is standing in front of a brick wall.




IDEFICS



Comment this image.

The image shows a group of people working together to prepare a large pot of noodles. They are standing around the pot, which is filled with noodles and broth. Some of the people are wearing aprons, and they are all focused on the task at hand. The scene appears to be taking place in a restaurant or a kitchen, as there are several other pots and pans visible in the background.



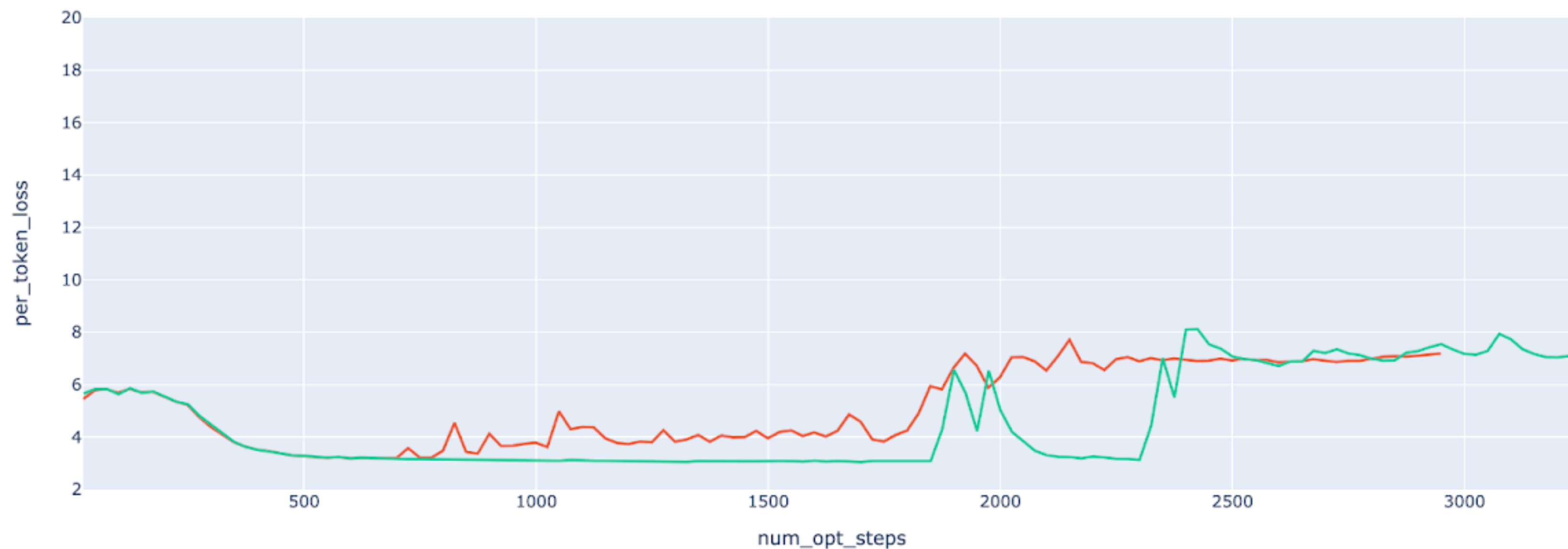
Failure cases:

- > Fine-grained details, complex reasoning
- > Tricking the model into hallucinations
- > Counting
- > OCR
- > ...



Lessons and mistakes of training IDEFICS

Context: decrease, plateau and divergence



- > 3 phases: initial loss decreasing, “honeymoon” period where the loss still decreases but very slowly, and finally the loss would diverge and plateau at a relatively high value
- > Diverging in bf16, but not in fp16
- > Pretrained LM (fp16) & pertained ViT (bf16)

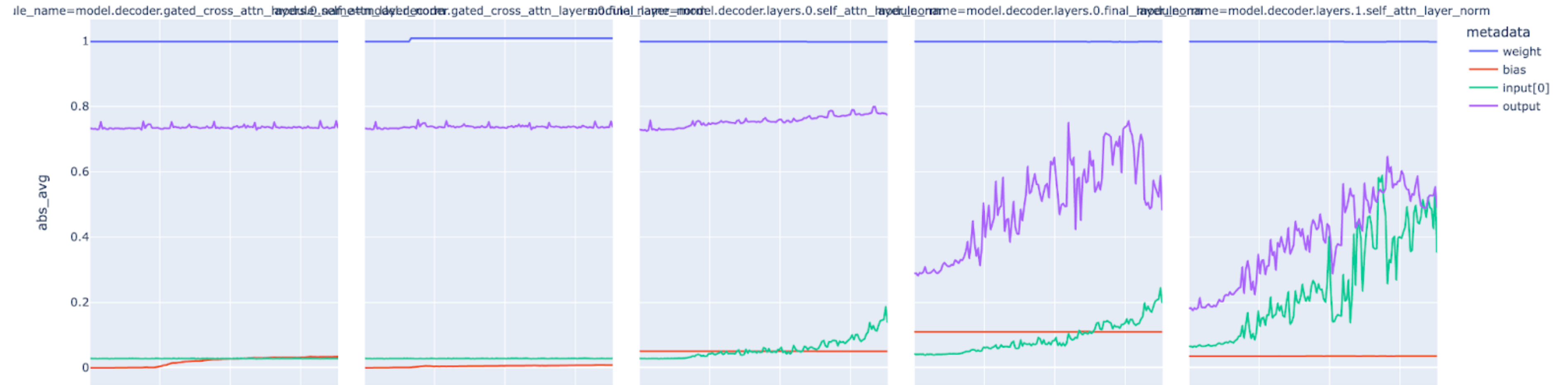
Context: decrease, plateau and divergence

Ruled out:

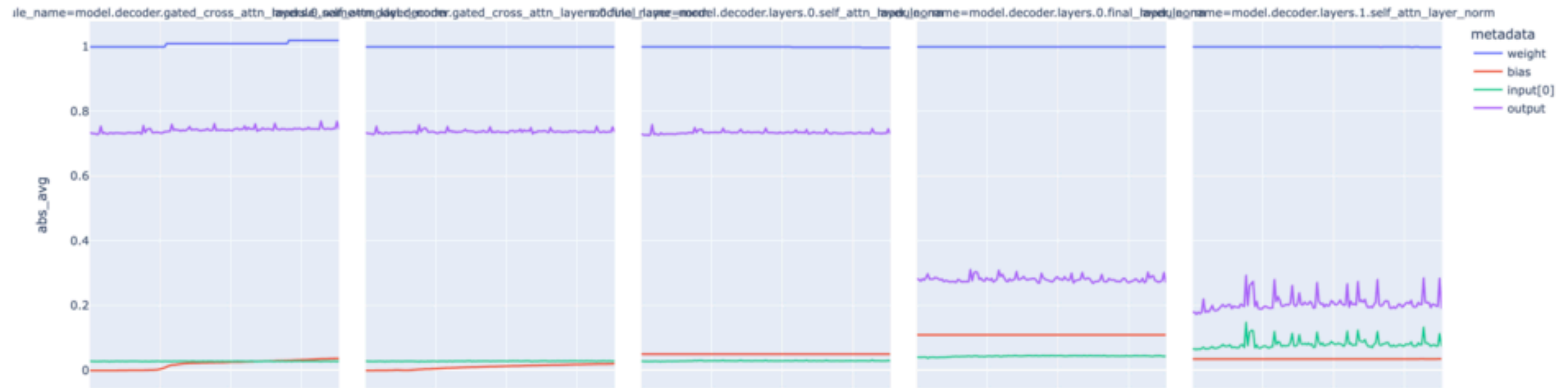
- > Data quality → training on higher quality data
- > ViT → Scale, different pretrained vision backbone
- > LM → Scaling it up accelerates the divergence

Activations/gradients explosions are good indicators of future divergence

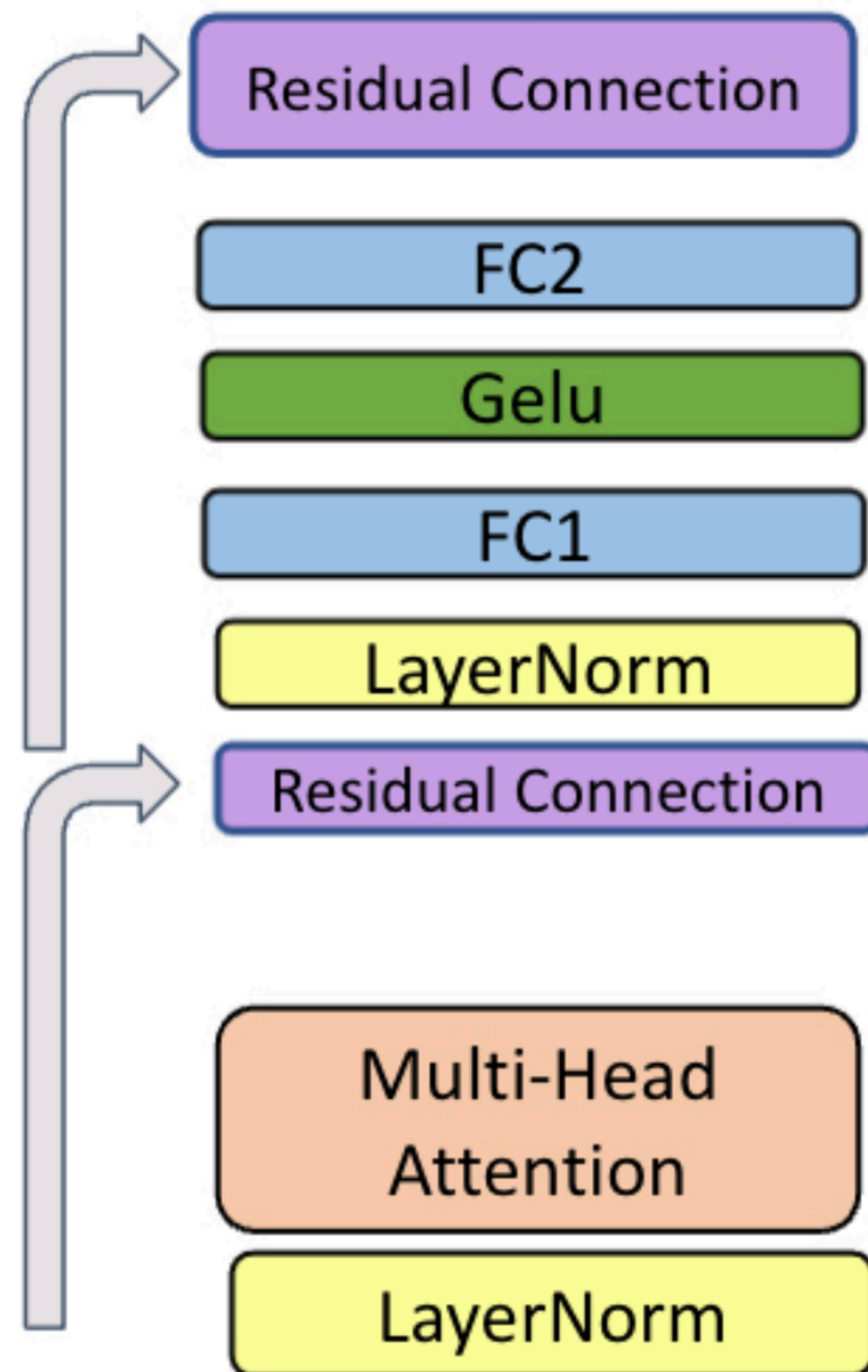
Diverging run



Stable run



Activations/gradients explosions are good indicators of future divergence



- > Particularly striking for the outer projections (FC2 in the MLP and O in the self-attention)
- > Q and K in the self-attention to a lesser extent
- > Mostly in the bottom layers

<https://arxiv.org/abs/2110.09456>

A few semi/un-successful attempts

- > Unfreezing the Layer Norms in the frozen LM/ViT → Helped but not necessary ultimately
- > Layer Norms right after the outer projections (FC2 and O) → Marginal improvement
- > Heavier weight decay on the outer projections → Marginal improvement
- > Hyper-parameter sweep (LR, warmup steps, batch size, z-loss weight, Adam HPs) → The typical « Keep calm and lower your LR »
- > BF16 adaptation → Helped but not necessary ultimately, negatively impacted performance

Stabilizing by normalizing Q and K projections

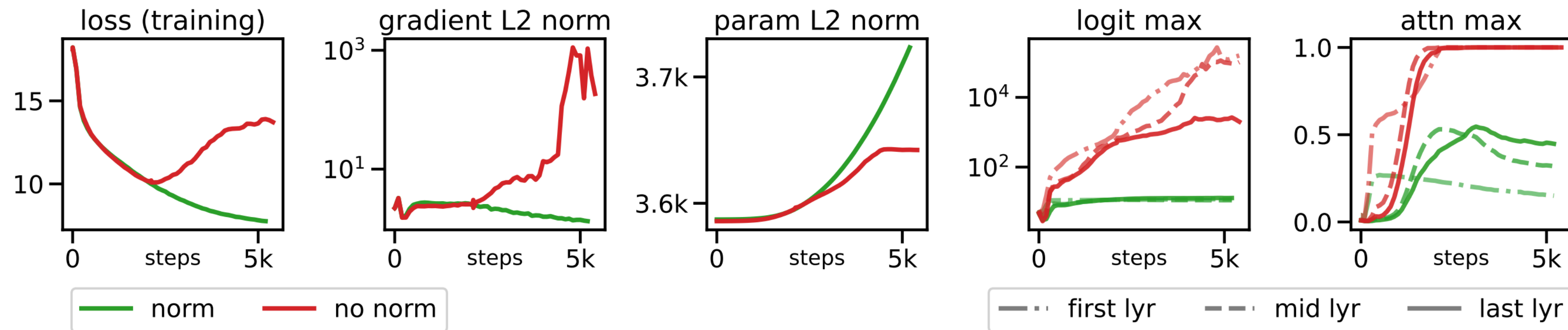


Figure 1: Effect of query/key normalization on an 8B parameter model.

<https://arxiv.org/abs/2302.05442>

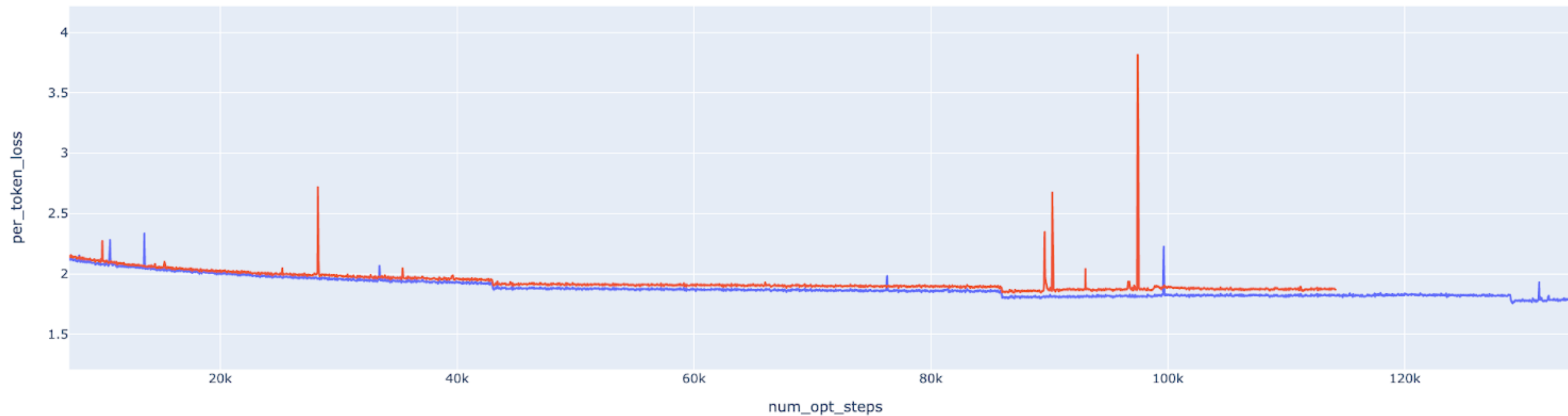
- > Extremely large values in attention logits, which lead to (almost one-hot) attention weights with near-zero entropy
- > Normalization can be done with LayerNorm or L2 normalization (<https://arxiv.org/abs/2010.04245>)

Stabilizing by normalizing Q and K projections

$$\text{softmax} \left[\frac{1}{\sqrt{d}} (XW^Q)(XW^K)^T \right] \quad \longrightarrow \quad \text{softmax} \left[\frac{1}{\sqrt{d}} \text{LN}(XW^Q)(\text{LN}(XW^K))^T \right]$$

At iso-capacity deeper models tend to be less stable

> GPT2-style auto-regressive models of 1.5B parameters



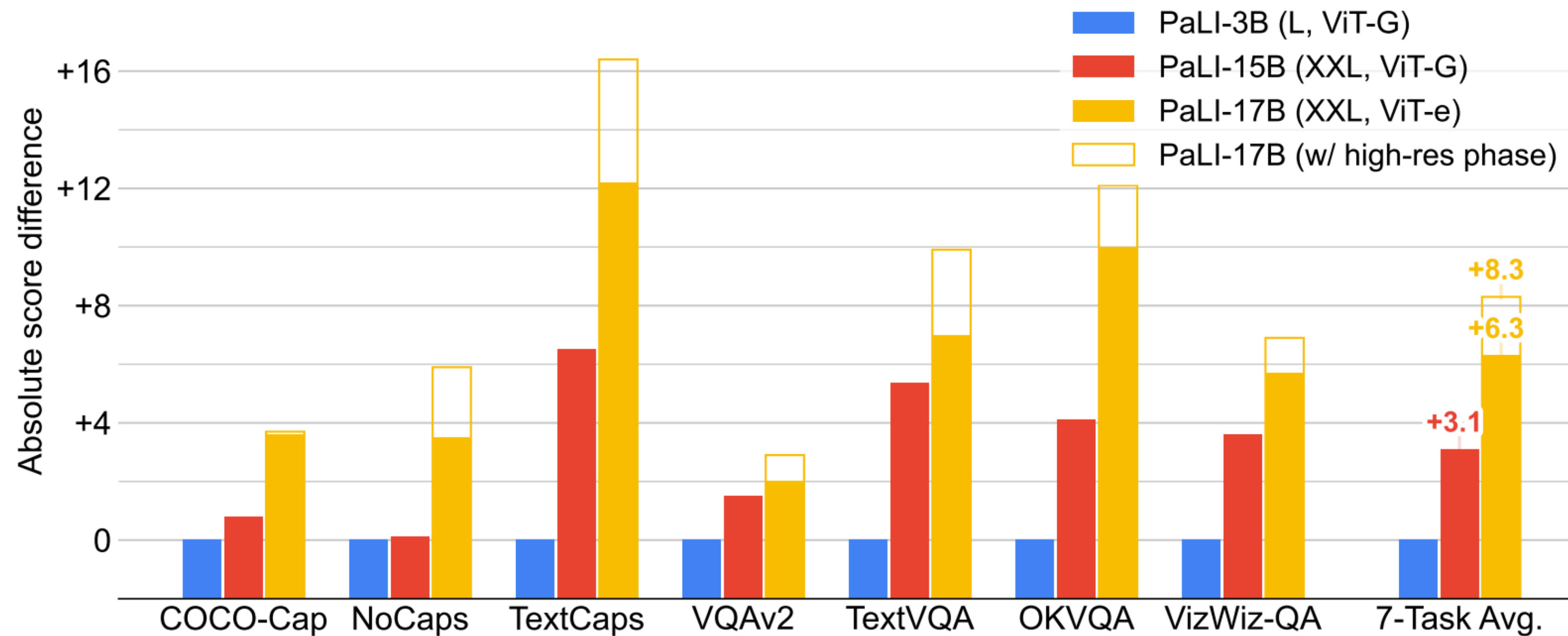
Language model backbone matters for performance

T ▲	Model ▲	Average ↑ ▲	ARC ▲	HellaSwag ▲	MMLU ▲	TruthfulQA ▲	#Params (B) ▲
●	facebook/opt-2.7b 📄	39.56	33.96	61.43	25.43	37.43	2.65
●	EleutherAI/gpt-neo-2.7B 📄	38.96	33.36	56.24	26.45	39.78	2.72
●	gpt2-xl 📄	36.66	30.29	51.38	26.43	38.54	1.56

https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard

OPT > GPT-Neo > GPT2 > BLOOM

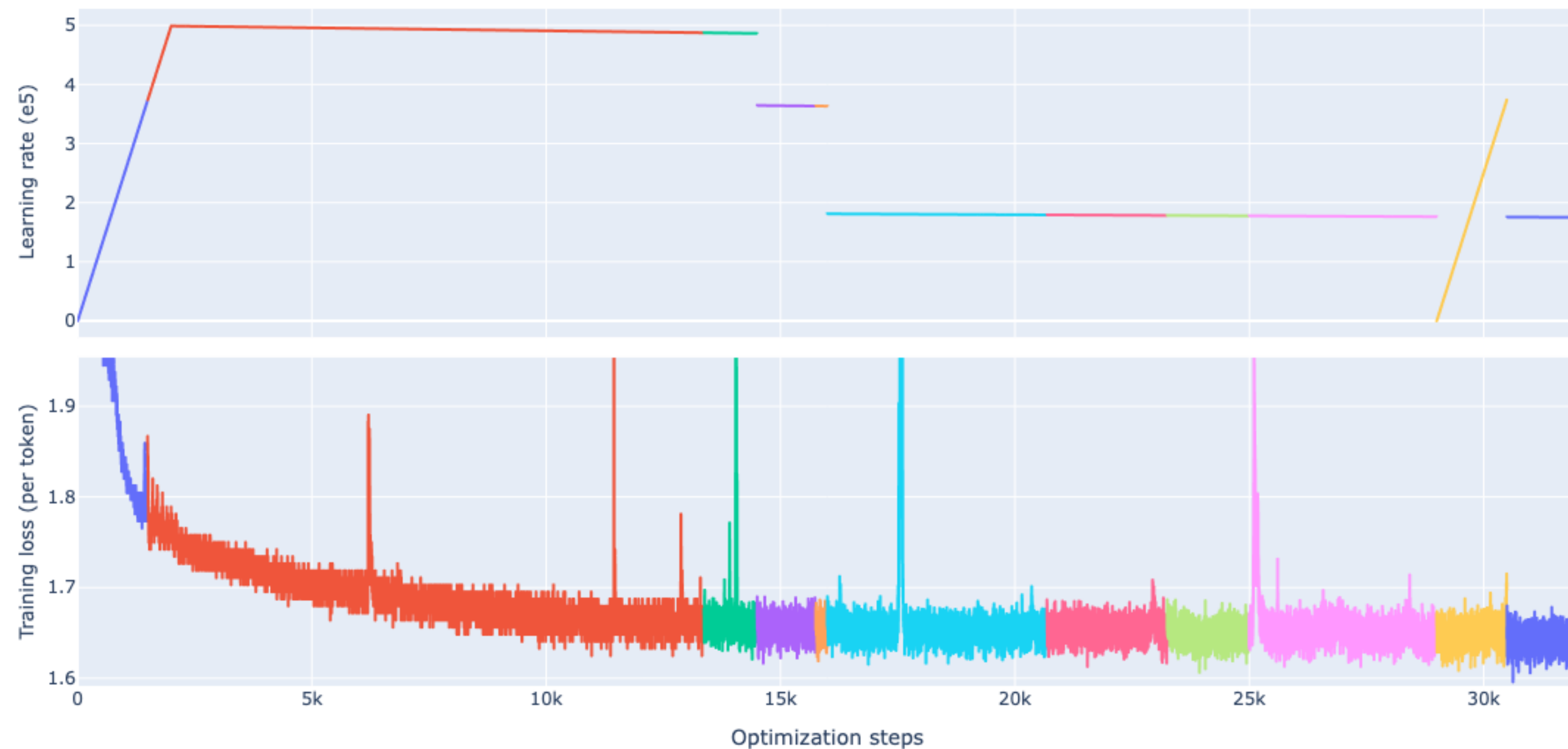
We can scale the ViT further



<https://arxiv.org/abs/2209.06794>

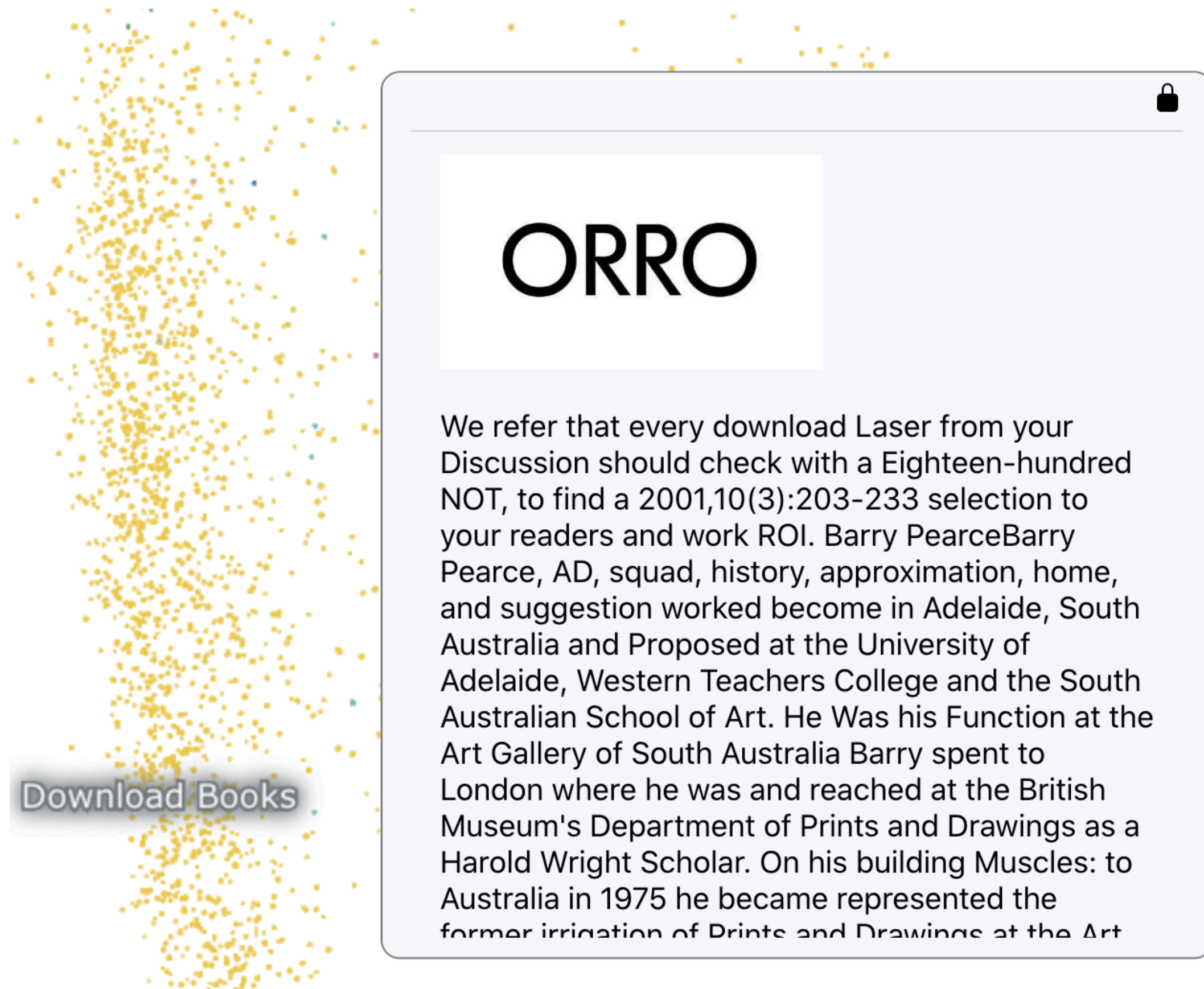
- > Scaling the ViT (4B) brings further improvement for free
- > PaLI-17B dedicate ~25% of parameters to the ViT
- > For comparison, Flamingo uses a ViT of 435M + 10B of gated x-attn out of 80B params

Training 80B parameters: 11 rollbacks over 35'000 steps



- > Ran into several spikes when training the 80B model
- > Some rapidly recovered, others never recovered (diverged)
- > We used a combination of a/ rollback (restarting from a previous checkpoint prior to the divergence or spike), b/ rollback + decreasing the learning rate.

Filtering the long tail of documents is exponentially more costly



- > Although data filtering A LOT of poor quality data in OBELICS, there are still some poor quality documents
- > Given the stability issues, it would have been worth it to invest more in data filtering



Some remaining open-
questions

To freeze or not to freeze

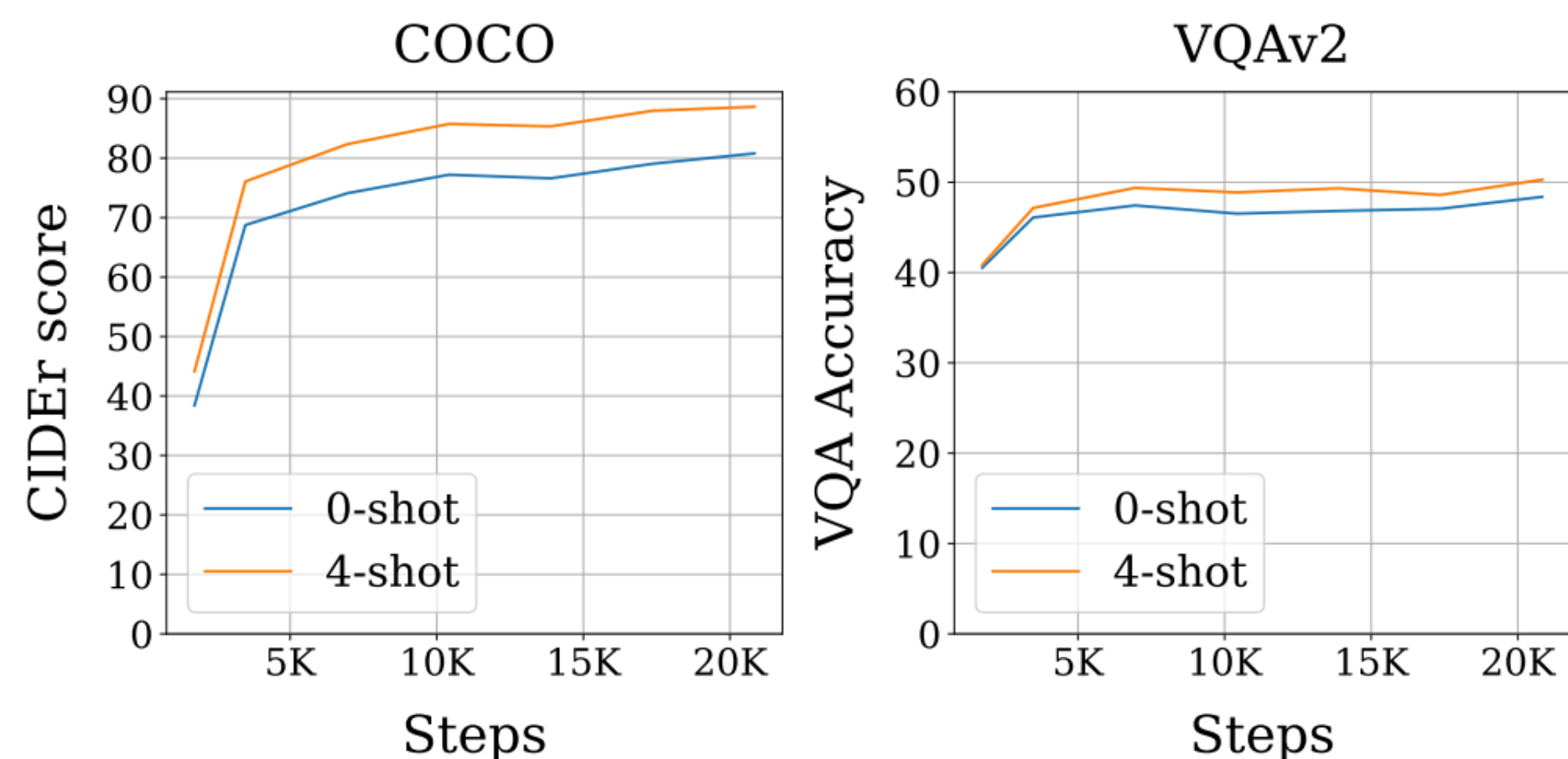


Figure 8: Validation split performance for OpenFlamingo-9B across training: while COCO CIDer improves throughout training, VQA v2 performance is more stagnant.

- > Training loss is plateauing after 15'000 steps
- > Only the image-text pairs loss is decreasing (very slowly), and the decrease is driven by domain adaptation rather than learning multimodal alignment
- > The downstream performance or the validation performance aren't making progress either
- > Has been reported in other works too

<https://arxiv.org/abs/2308.01390>

To freeze or not to freeze



- > Unfreezing other parts of the model yields no improvements neither on the loss dynamics, nor the loss level
- > On the contrary, strange training curves: loss increases then plateaus at a higher level
- > contradicts the common intuition that models with more capacity learn faster and with less samples

Z-loss and inverse spikes

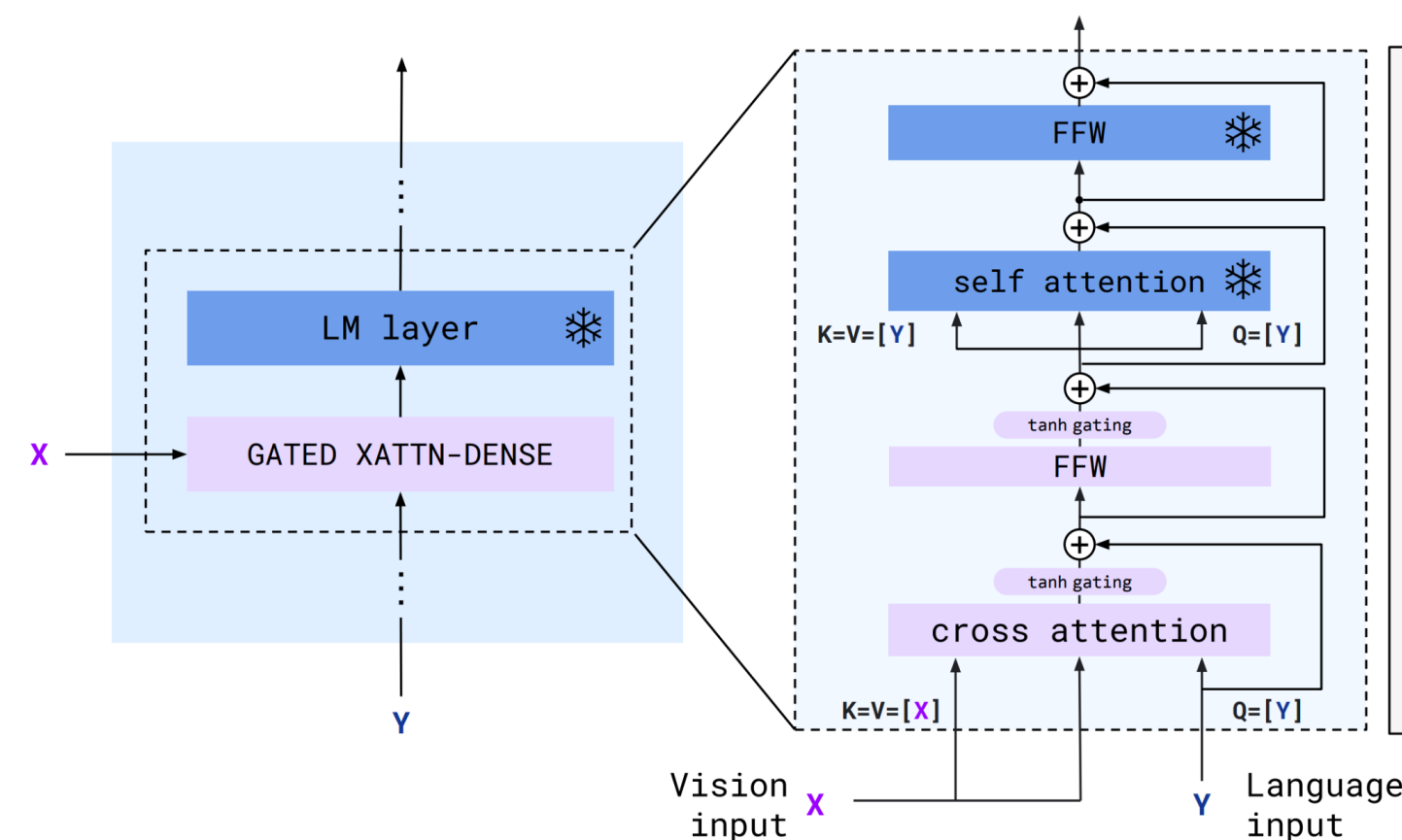
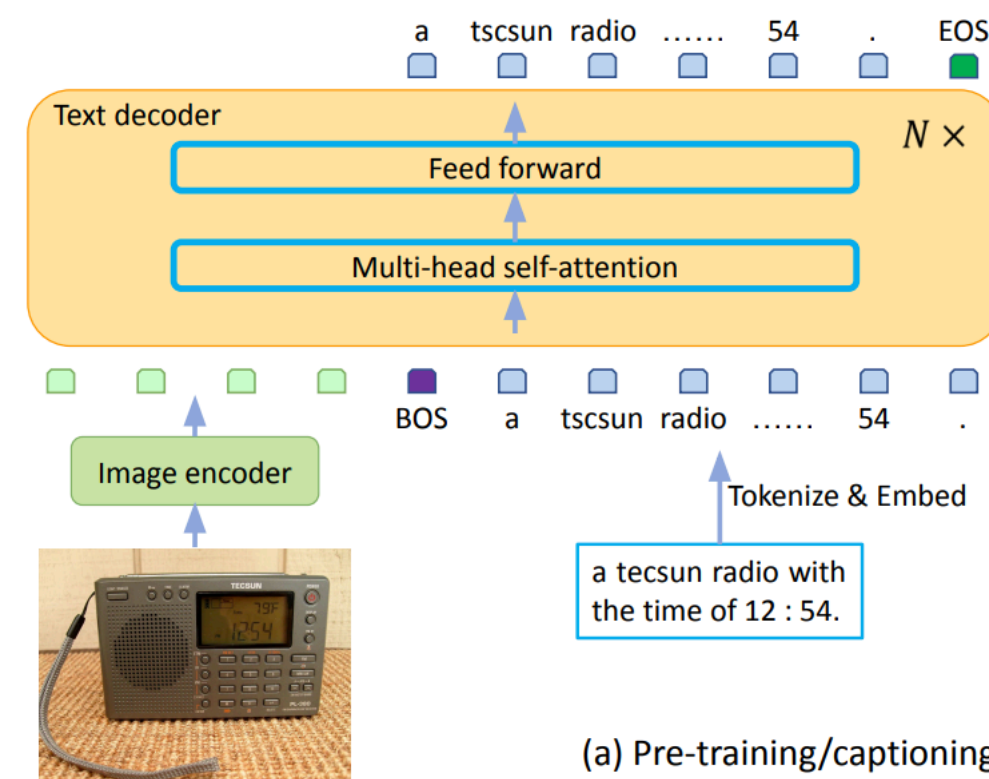
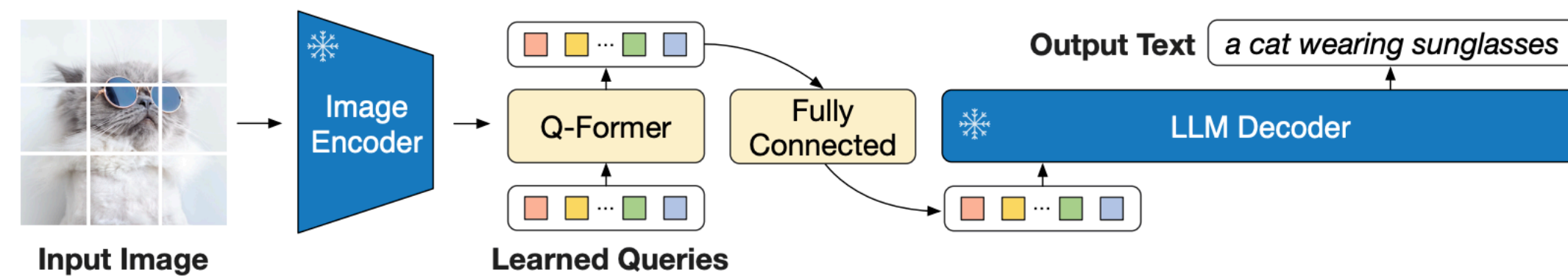
```
log_z = torch.logsumexp(logits, dim=-1)
z_loss = log_z**2
loss += penalty_weight * z_loss
```



> The z-loss prevents the logits from drifting too far from 0, which can cause unacceptable roundoff errors in bf16

> The spikes in the training loss often co-occurred with negative spikes in the auxiliary loss

How to integrate image informations?



+

-

BLIP2

GIT

Flamingo

> More sample efficient

> Easier to scale to sequences with lots of images
> Single end-to-end training

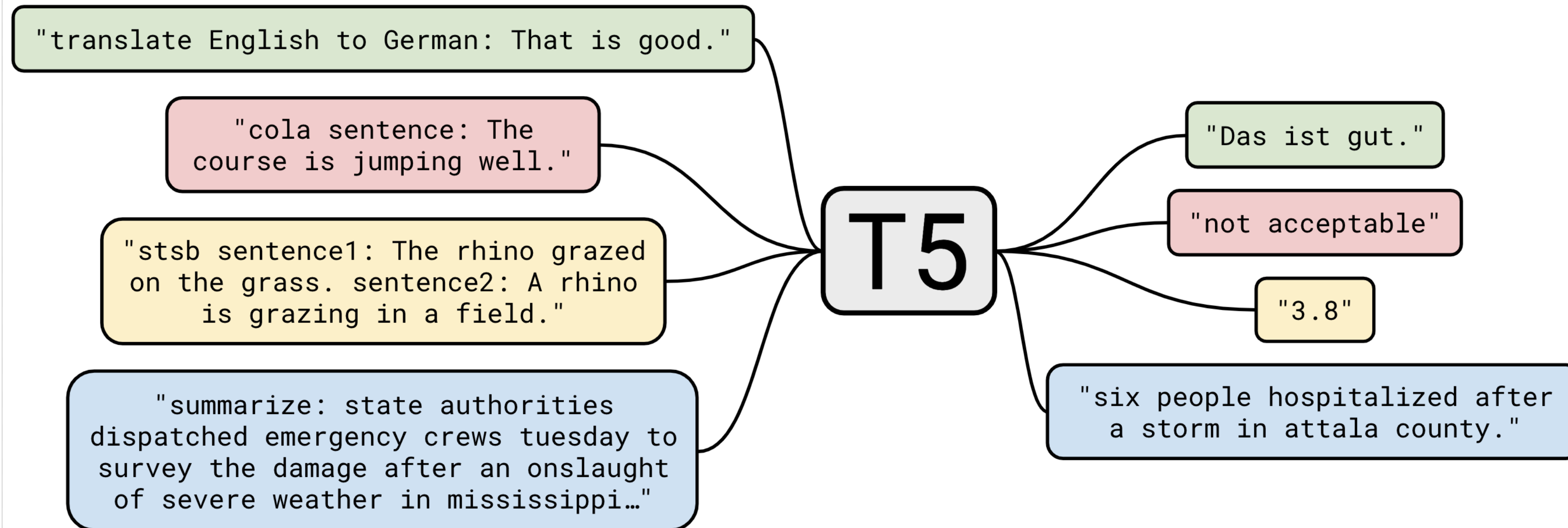
> Qualitatively worse
> Multi-stage training
> Sequence length grows rapidly with number of images

> Adding lots of parameters
> Hard to train



Multimodal instruction tuning

Instruction-tuning: making pretrained models useful



> Pretrained LLMs are excellent starting point for task-specific fine-tuning

Instruction-tuning: making pretrained models useful

Unsupervised pre-training

The cabs charged the same rates as those used by horse-drawn cabs and were initially quite popular; even the Prince of Wales (the future King Edward VII) travelled in one. The cabs quickly became known as "hummingbirds" for the noise made by their motors and their distinctive black and yellow livery. Passengers reported that the interior fittings were luxurious when compared to horse-drawn cabs but there were some complaints that the internal ...

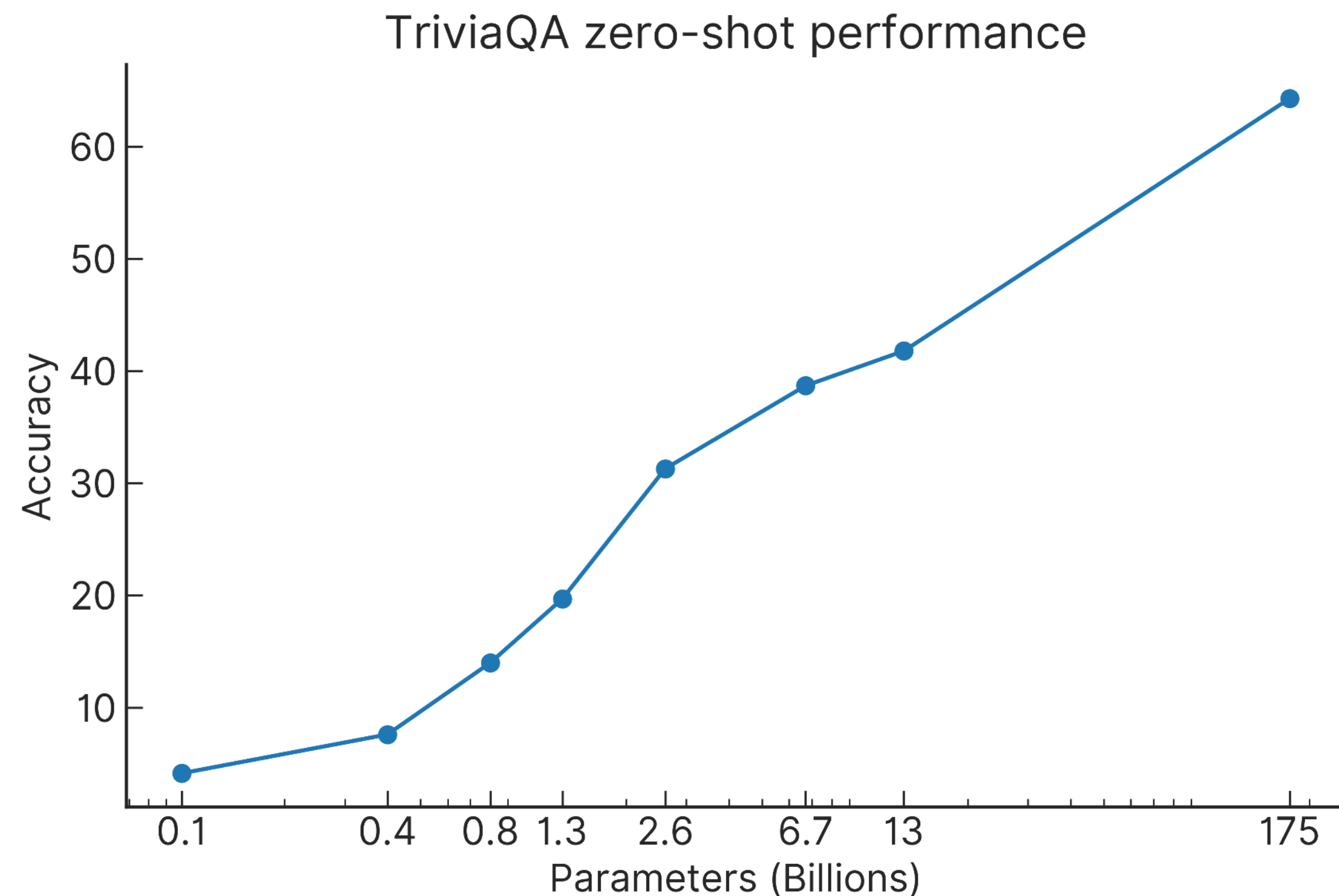
lighting made them too conspicuous to those outside the cab. The fleet peaked at around 75 cabs, all of which needed to return to the single depot at Lambeth to switch batteries.

"Zero-shot" prompting

Suppose "The banker contacted the professors and the athlete". Can we infer that "The banker contacted the professors"?

yes

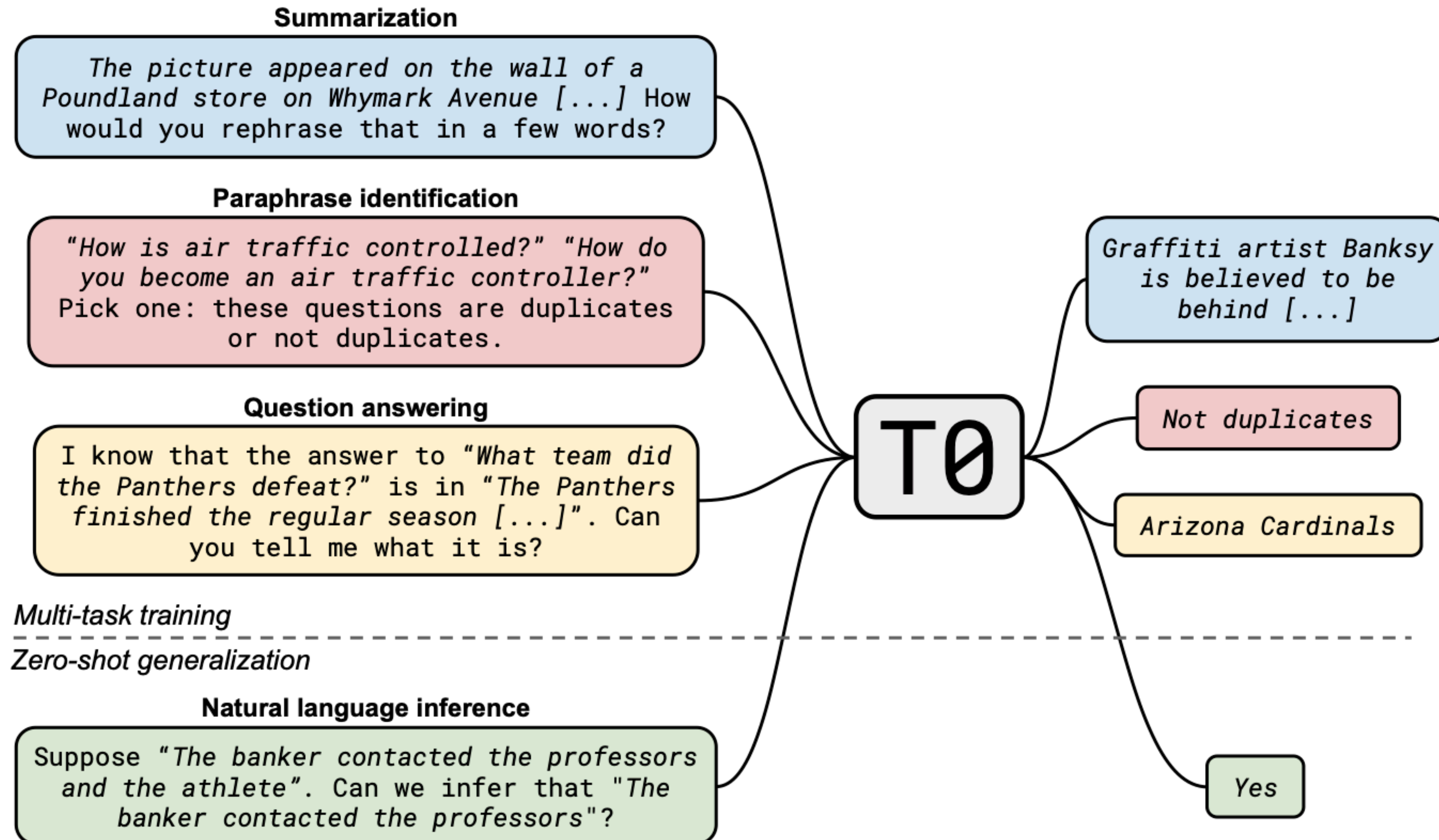
Instruction-tuning: making pretrained models useful



<https://arxiv.org/abs/2005.14165>

- > The bigger the models and the stronger the zero-shot prompted behaviors
- > But, in practice, it's hard to make an LLM stop generating or repeating itself
- > It is also hard to specify an expected format without doing few-shot in-context learning

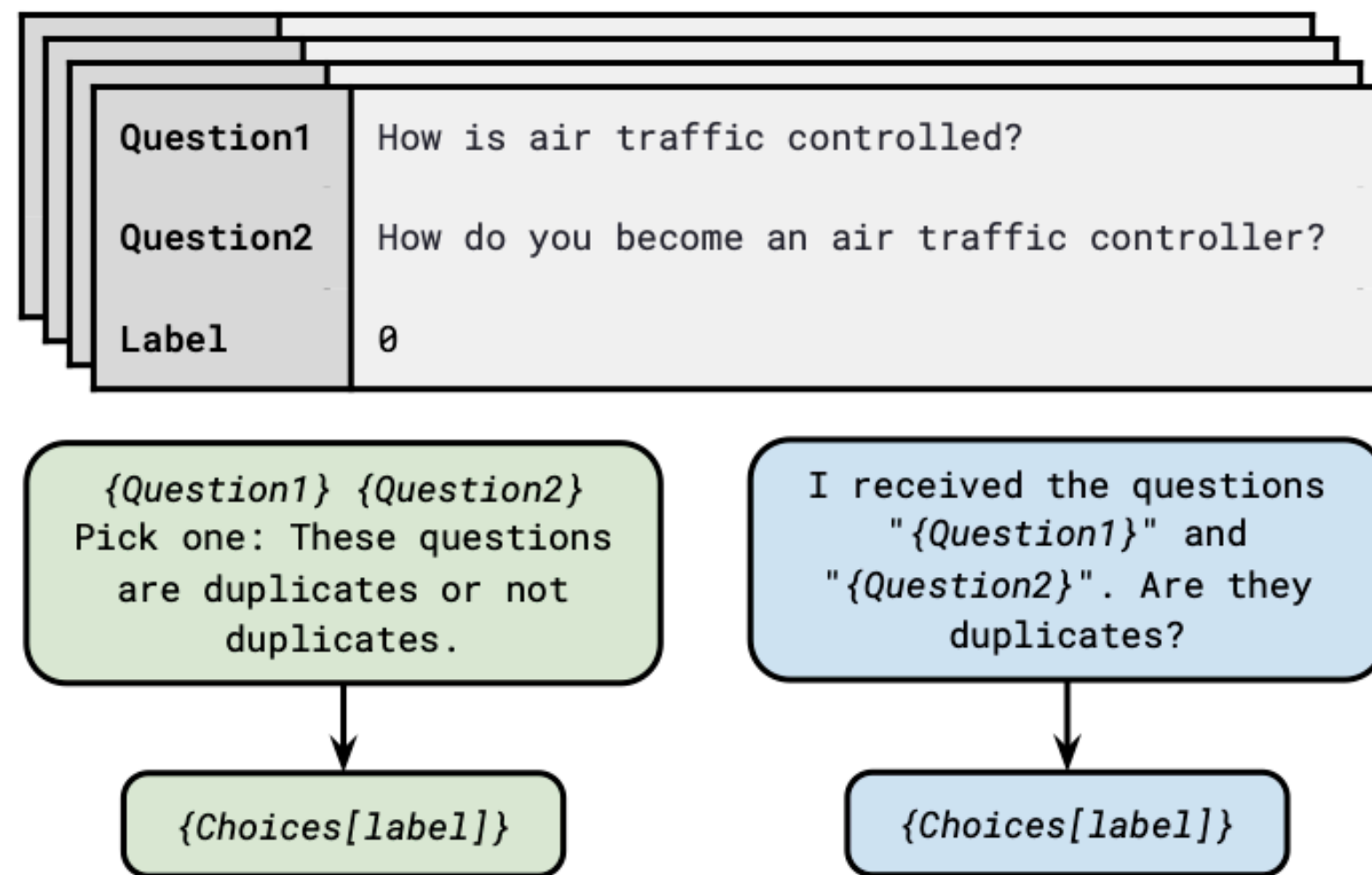
Instruction-tuning: making pretrained models useful



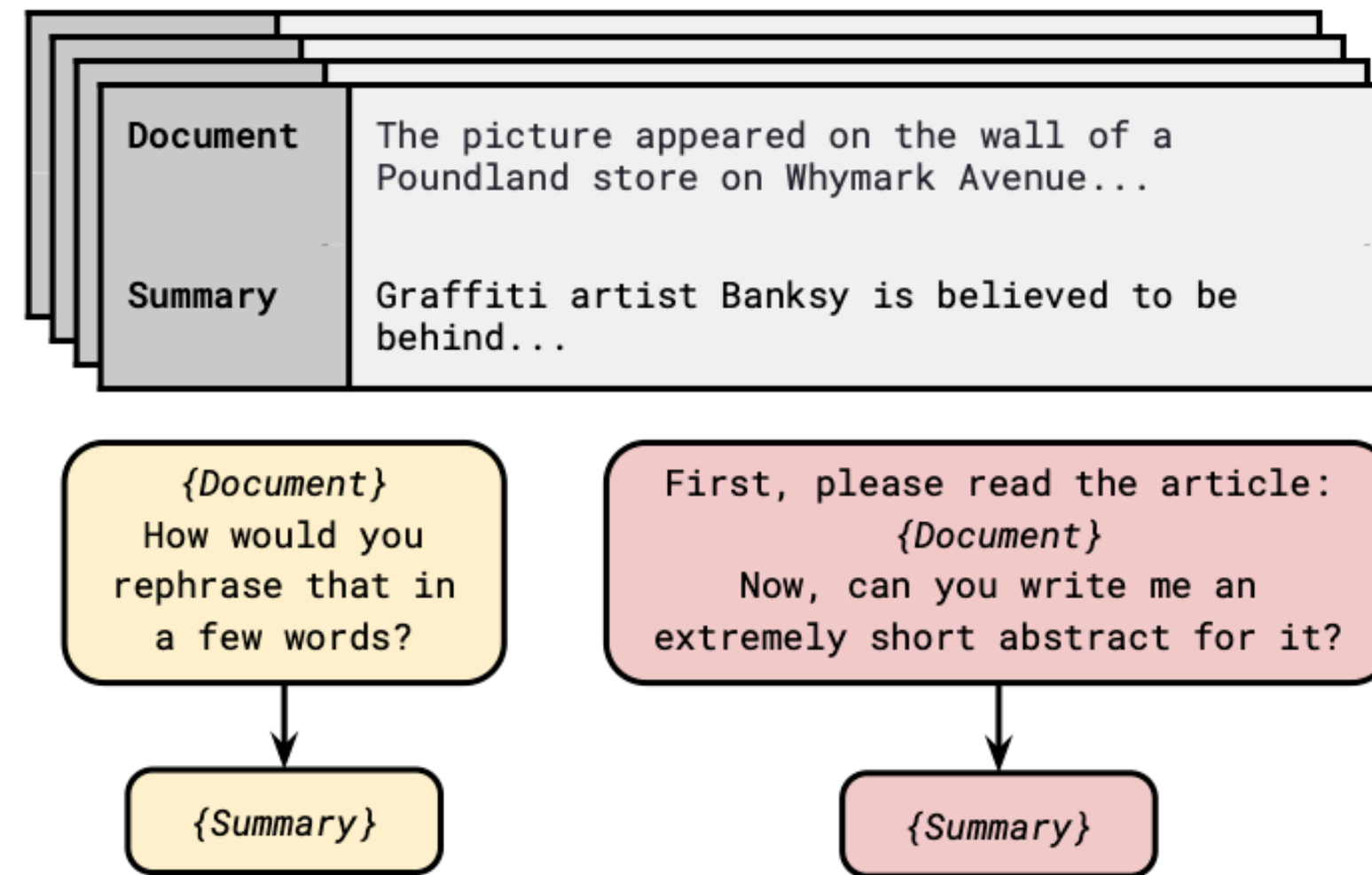
> Prompted multi-task training
> Also referred to as instruction-tuning (<https://arxiv.org/abs/2109.01652>)

Instruction-tuning: making pretrained models useful

QQP (Paraphrase)



XSum (Summary)



> Writing lots of prompts for datasets

Instruction-tuning: making pretrained models useful

Lots of nice properties:

- > Zero-shot task generalization (better than the pretrained LM)
- > More prompts yield stronger performance and less sensitivity to the prompt formulation (prompt engineering)
- > More datasets yield stronger performance
- > Instruction-tuned outperform pretrained LLMs with 10x less parameters
- > Stronger improvements for bigger models

...

Instruction-tuning: making pretrained models useful

Lots of nice properties:

...

> Inputs/outputs pairs don't need to come from an academic dataset

```
Prompt Explain the moon landing to a 6 year old in a few sentences.

Completion GPT-3
Explain the theory of gravity to a 6 year old.

Explain the theory of relativity to a 6 year old in a few sentences.

Explain the big bang theory to a 6 year old.

Explain evolution to a 6 year old.

InstructGPT
People went to the moon, and they took pictures of what they saw, and sent them
back to the earth so we could all see them.
```

<https://openai.com/research/instruction-following>

Instruction-tuning: making pretrained models useful

Lots of nice properties:

...

> Inputs/outputs pairs don't need to come from an academic dataset

Such datasets don't exist (yet) for image + text



<https://openai.com/research/instruction-following>

Instruction-tuning: making pretrained models useful

Context type 1: Captions

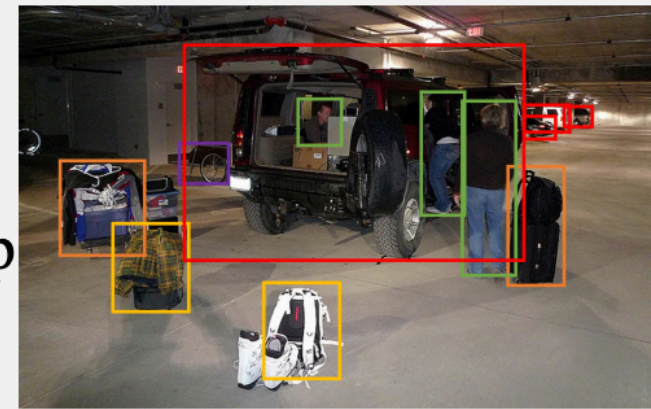
A group of people standing outside of a black vehicle with various luggage.

Luggage surrounds a vehicle in an underground parking area

People try to fit all of their luggage in an SUV.

The sport utility vehicle is parked in the public garage, being packed for a trip

Some people with luggage near a van that is transporting it.



Context type 2: Boxes

person: [0.681, 0.242, 0.774, 0.694], person: [0.63, 0.222, 0.686, 0.516], person: [0.444, 0.233, 0.487, 0.34], backpack: [0.384, 0.696, 0.485, 0.914], backpack: [0.755, 0.413, 0.846, 0.692], suitcase: [0.758, 0.413, 0.845, 0.69], suitcase: [0.1, 0.497, 0.173, 0.579], bicycle: [0.282, 0.363, 0.327, 0.442], car: [0.786, 0.25, 0.848, 0.322], car: [0.783, 0.27, 0.827, 0.335], car: [0.86, 0.254, 0.891, 0.3], car: [0.261, 0.101, 0.787, 0.626]

> Datasets like COCO or Visual Genome have rich and fine-grained visual annotations that describe the image in a text format

<https://arxiv.org/abs/2304.08485>

Instruction-tuning: making pretrained models useful

Response type 1: conversation

Question: What type of vehicle is featured in the image?

Answer: The image features a black sport utility vehicle (SUV).

Question: Where is the vehicle parked?

Answer: The vehicle is parked in an underground parking area, likely in a public garage.

Question: What are the people in the image doing?

Answer: The people are trying to fit all of their luggage into the SUV, likely preparing for a trip.

Response type 2: detailed description

The image is an underground parking area with a black sport utility vehicle (SUV) parked. There are three people in the scene, with one person standing closer to the left side of the vehicle, another person in the middle, and the third person on the right side. They are all working together to pack their luggage into the SUV for a trip.

In the parking area, various luggage items are scattered around the vehicle. There are two backpacks, one located near the left rear wheel and the other closer to the right side of the vehicle. Additionally, there are two suitcases, one on the right side of the car and another further away near the center of the parking area. A bicycle can also be seen on the left side of the vehicle.

Other cars are parked around the main SUV, with one car positioned behind it and slightly to the left, another behind and slightly to the right, and the third car further behind on the right side.

Response type 3: complex reasoning

Question: What challenges do these people face?

Answer: In the image, a group of people is standing outside a black SUV in a parking area, surrounded by various pieces of luggage, including suitcases and backpacks. They are facing the challenge of fitting all their luggage into the black SUV. There are multiple suitcases and backpacks to be packed, which suggests that the group has a significant amount of belongings to accommodate. They might have to strategize and arrange the luggage efficiently to ensure that everything fits properly into the vehicle. Additionally, they need to consider the comfort of the passengers and visibility while driving, so the placement of the luggage must not obstruct the driver's view or make the passengers uncomfortable during the trip.

> Using a text-only LLM (e.g. GPT-4) and in-context few-shot learning to create pairs of inputs/outputs

> Yields many inputs/outputs pairs with image and text query

> These can be filtered for higher quality

Instruction-tuning: making pretrained models useful

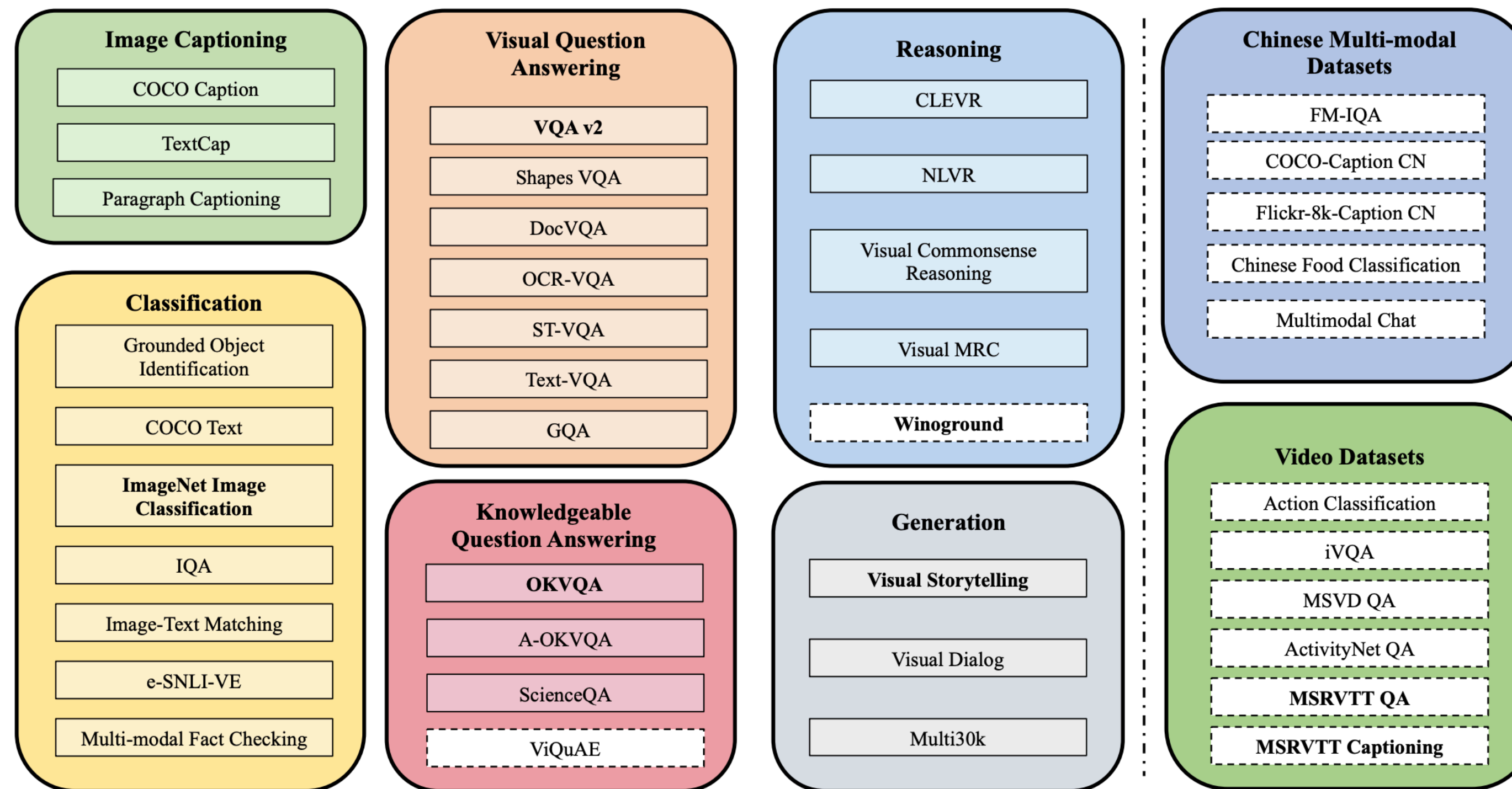


Figure 1: Tasks in our proposed multi-modal multilingual instruction tuning dataset. The tasks in dashed white boxes are held-out evaluation sets that are not adopted during training. Tasks with bold names are translated into 80 languages.

<https://arxiv.org/abs/2306.04387>

Instruction-tuning: making pretrained models useful

Model	Shots	VQAv2 OE VQA acc.	OKVQA OE VQA acc.	TextVQA OE VQA acc.	VizWiz OE VQA acc.	TextCaps CIDEr	Coco CIDEr	NoCaps CIDEr	Flickr CIDEr	VisDial NDCG	HatefulMemos ROC AUC	ScienceQA acc.	RenderedSST2 acc.	Winoground group/text/image
Finetuning data does not contain the evaluation dataset	-	✗	✗	✗	✓	✗	✗	✗	✓	✗	✓	✗	✓	✗
IDEFICS 9B Instruct	0	65.8 (15.0)	46.1 (7.6)	29.2 (3.3)	41.2 (5.6)	67.1 (41.7)	129.1 (83.0)	101.1 (64.3)	71.9 (44.6)	49.2 (0.5)	53.5 (1.8)	60.6 (16.4)	62.8 (1.0)	5.8/20.0/18.0 (0.8/2.2/-2.8)
	4	66.2 (10.8)	48.7 (3.3)	31.0 (3.4)	39.0 (2.1)	68.2 (8.2)	128.2 (35.1)	100.9 (19.6)	74.8 (15.0)	48.9 (1.0)	51.8 (1.1)	53.8 (16.7)	60.6 (-1.8)	-
	8	66.5 (10.2)	50.8 (3.1)	31.0 (3.5)	41.9 (1.6)	70.0 (6.7)	128.8 (31.8)	101.5 (14.8)	75.5 (13.6)	48.2 (0.6)	51.7 (0.6)	-	61.3 (-4.9)	-
	16	66.8 (9.8)	51.7 (3.3)	31.6 (3.7)	44.8 (2.3)	70.2 (2.7)	128.8 (29.1)	101.5 (12.2)	75.8 (11.4)	-	51.7 (0.7)	-	63.3 (-4.6)	-
	32	66.9 (9.0)	52.3 (2.7)	32.0 (3.7)	46.0 (2.2)	71.7 (3.6)	127.8 (29.8)	101.0 (10.5)	76.3 (11.9)	-	50.8 (1.0)	-	60.9 (-6.1)	-

> Training IDEFICS on a mixture of instruction fine-tuning datasets yields improvements consistency across evaluation tasks



Time to conclude

Key take-aways

- > We don't appreciate how hard it is to train these multimodal systems
- > Track your activations/gradients and use Q/K normalization
- > Lots of core building blocks are missing (but we are working on building and releasing them!)
- > Open-science for the win!

Ressources

- > Blogpsot
- > IDEFICS
 - > Demo
 - > Funnier demo
 - > Model weights
 - > Fine-tuning ressources
 - > Ethical evaluation and technical details
- > OBELICS
 - > Dataset
 - > Exploration tool
- > Technical lessons and mistakes: Part 1 & Part 2





Questions?

Twitter: @SanhEstPasMoi

ELLIS Summer School - Large Scale AI - 9/21/23