

AI research:

The unreasonably narrow path and how not to be miserable

Google Tech Talk

21 Oct 2020

Rosanne Liu

<https://rosanneliu.com>

This is a true story

This is a true story

Long, long time ago,

This is a true story

Long, long time ago, in February 2020,

This is a true story

Long, long time ago, in February 2020,
I was laid off from Uber AI

This is a true story

Long, long time ago, in February 2020,
I was laid off from Uber AI

++: The whole AI Labs (the research arm) were laid off

This is a true story

Long, long time ago, in February 2020,
I was laid off from Uber AI

++: The whole AI Labs (the research arm) were laid off

--: I was part of the founding team

This is a true story

Long, long time ago, in February 2020,
I was laid off from Uber AI

++: The whole AI Labs (the research arm) were laid off

--: I was part of the founding team

++: “I am an AI researcher!”

This is a true story

Long, long time ago, in February 2020,
I was laid off from Uber AI

++: The whole AI Labs (the research arm) were laid off

--: I was part of the founding team

++: “I am an AI researcher!”

--: A pandemic is looming

Anyway, I started job hunting

Anyway, I started job hunting

And that's when it becomes clear how woefully
narrow a path we have in front of us,

Anyway, I started job hunting

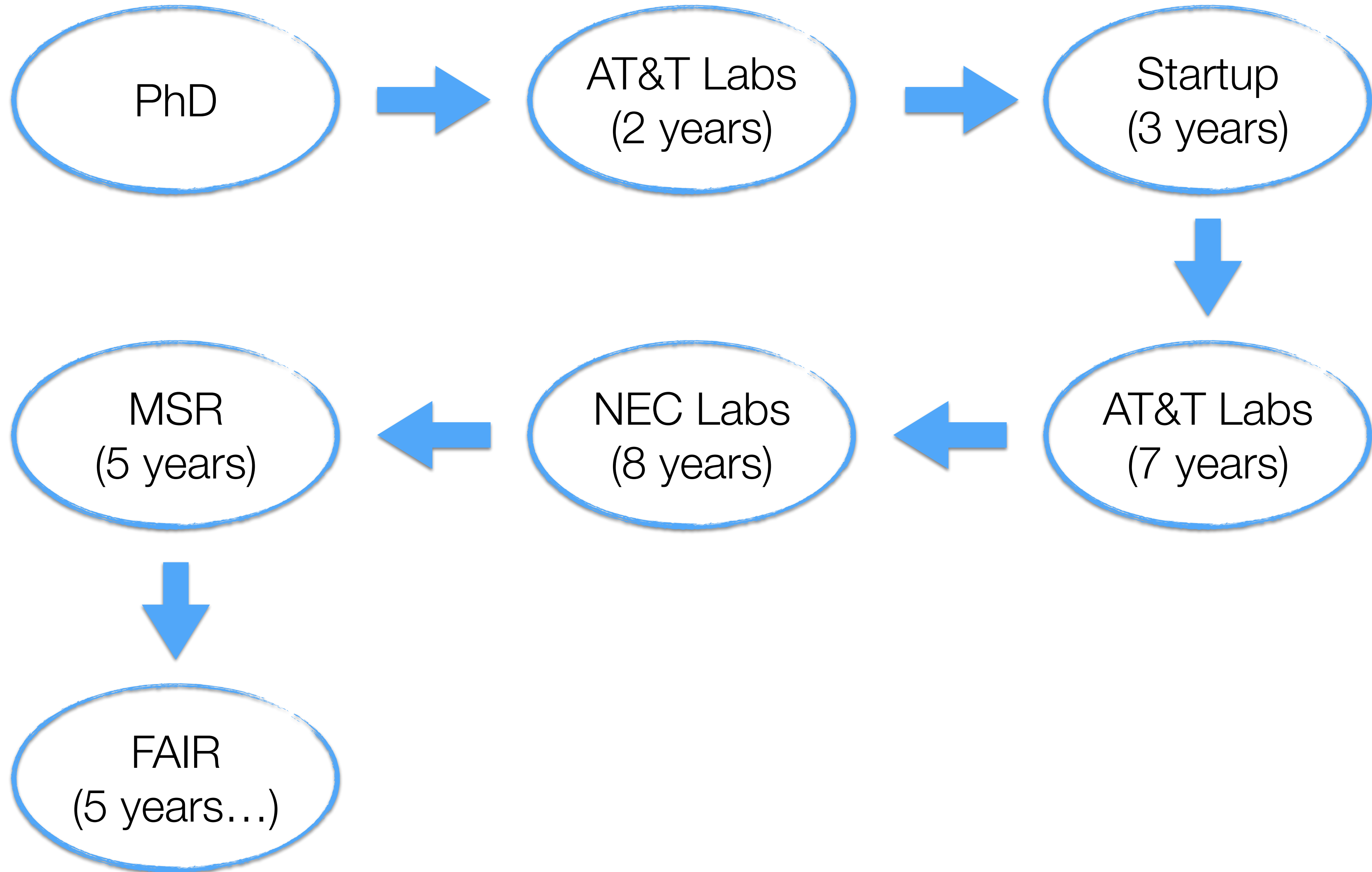
And that's when it becomes clear how woefully
narrow a path we have in front of us,

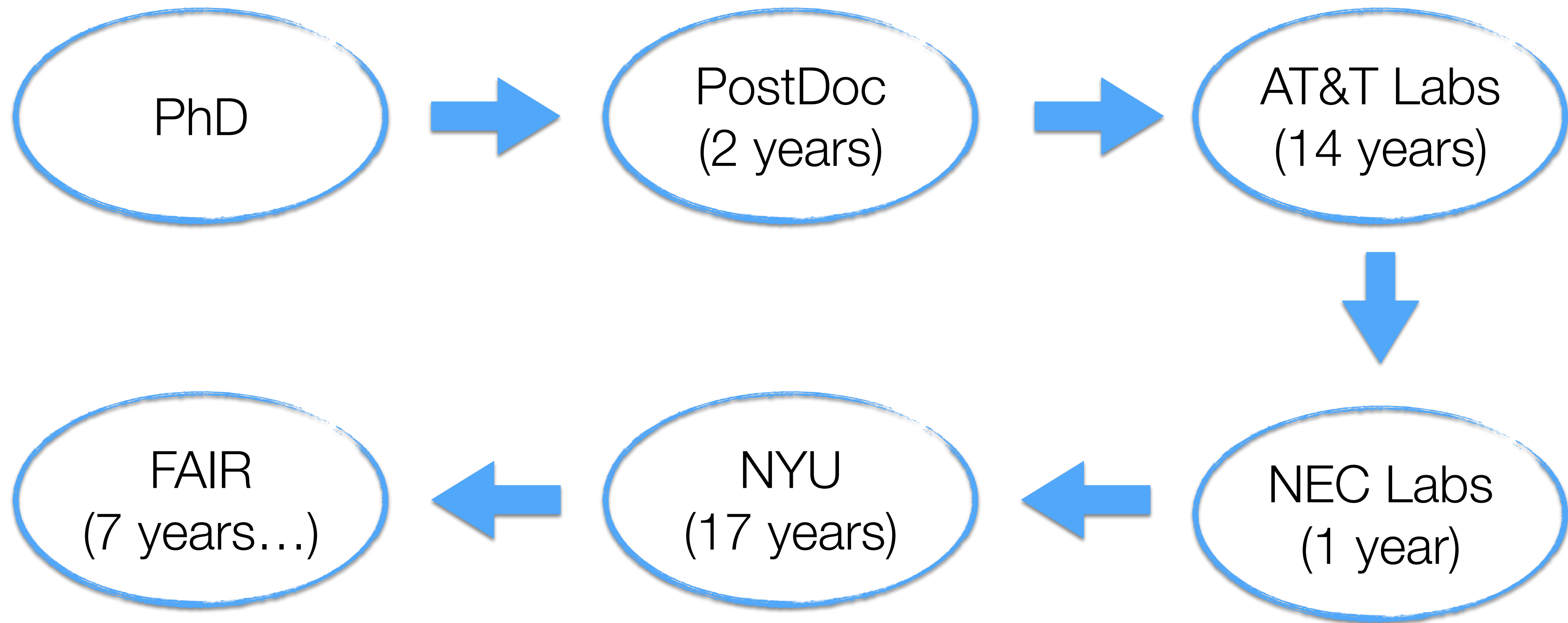
and how that, in turn, has made me
narrow-minded (and miserable).

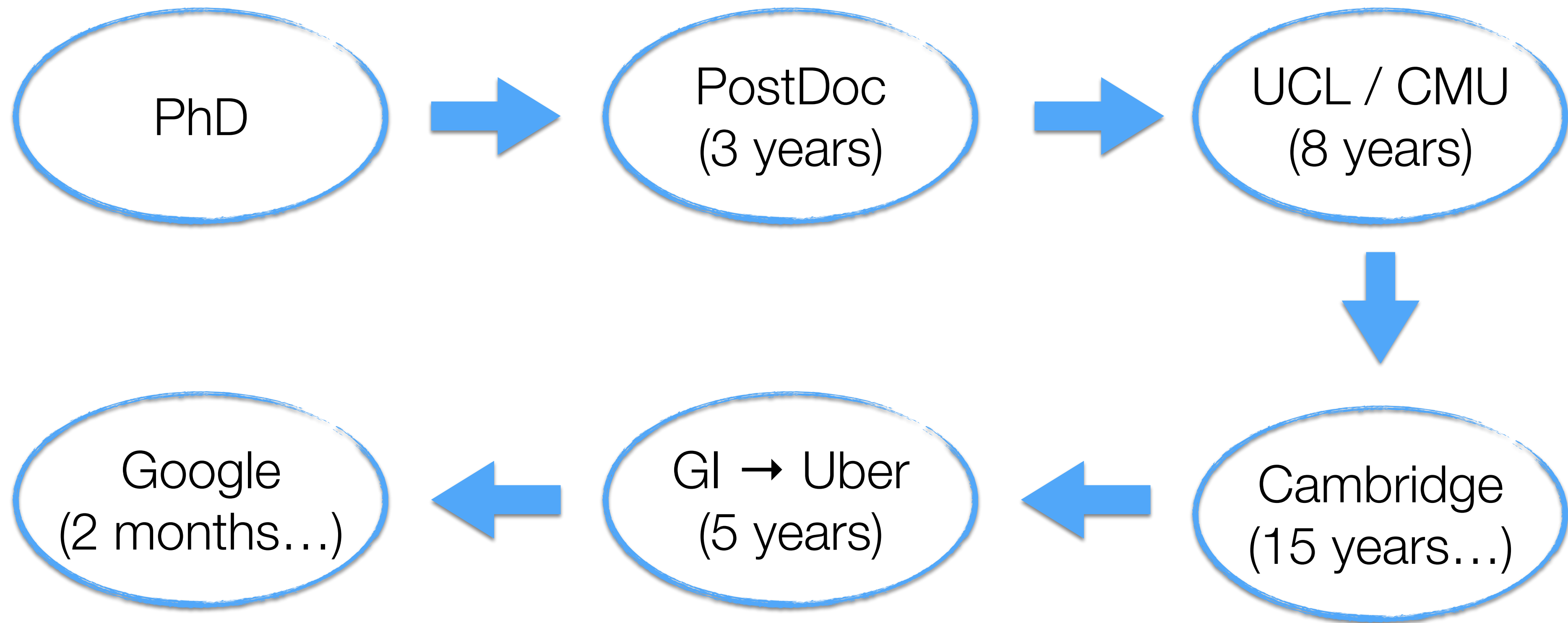
Anyway, I started job hunting

And that's when it becomes clear how woefully
narrow a path we have in front of us,

and how that, in turn, has made me
narrow-minded (and miserable).







Apparently there are two paths for AI research

- Academia, often in one place
- Industry, likely on cycles

Apparently there are two paths for AI research

- Academia, often in one place
- Industry, likely on cycles

But that's not why I call it *"narrow"*

Apparently there are two paths for AI research

- Academia, often in one place
- Industry, likely on cycles

It is narrow when... all of them are trying to hire the
same kind of people, with the same rigid rubric.

Pain point #1: Rubrics for hiring are highly correlated across all major industrial labs.

Similar to papers published at all the conferences?

To give you an idea how I don't fit the rubric...

- *“You were the middle author in a lot of papers, what exactly were your contributions?”*
- *“Almost all of your papers have Jason on it, who are you removed from him?”*
- *“I can't quite place you as an expert of anything; what do you want to focus on next?”*

- *“You were the middle author in a lot of papers, what exactly were your contributions?”*

Pain point

Reason

- *“You were the middle author in a lot of papers, what exactly were your contributions?”*

Pain point

The increasingly messy credit assignment problem.

Reason

- *“You were the middle author in a lot of papers, what exactly were your contributions?”*

Pain point

The increasingly messy credit assignment problem.

Reason

There's now too much (quantifiable) gains associated with ML papers.

- *“Almost all of your papers have Jason on it, who are you removed from him?”*

Pain point

Reason

- *“Almost all of your papers have Jason on it, who are you removed from him?”*

Pain point

Behind any mildly successful woman there is a white man. (No sarcasm. It's true!)

Reason

- *“Almost all of your papers have Jason on it, who are you removed from him?”*

Pain point

Behind any mildly successful woman there is a white man. (No sarcasm. It's true!)

Reason

Stereotype?

- *“I can’t quite place you as an expert of anything; what do you want to focus on next?”*

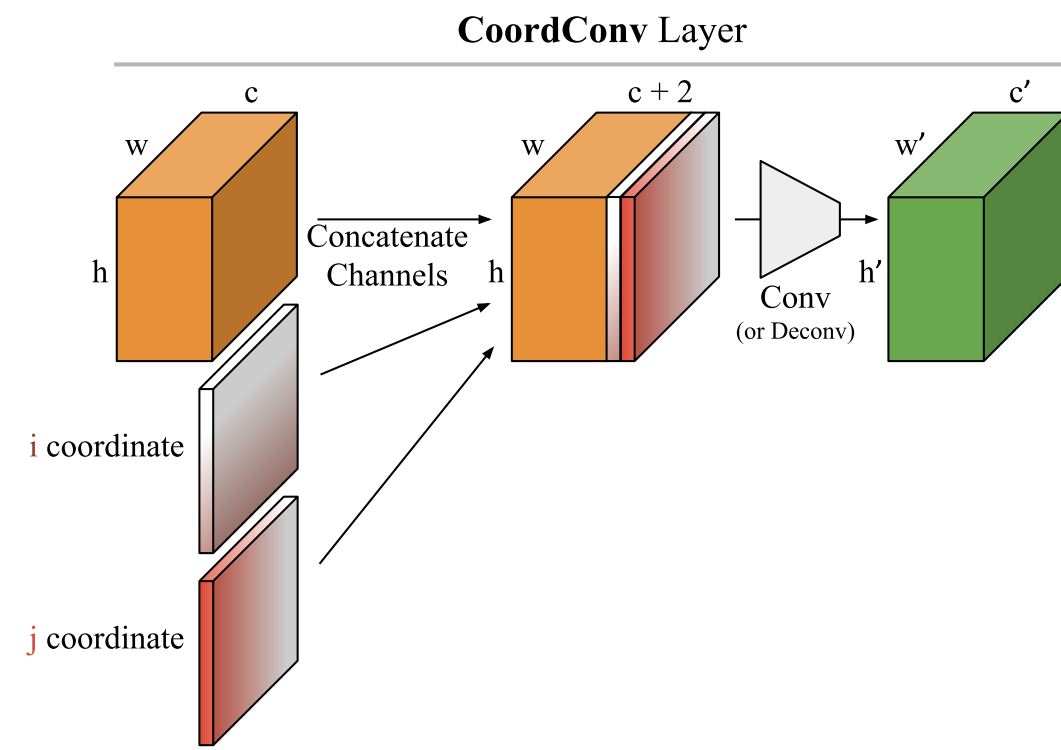
Pain point

Reason

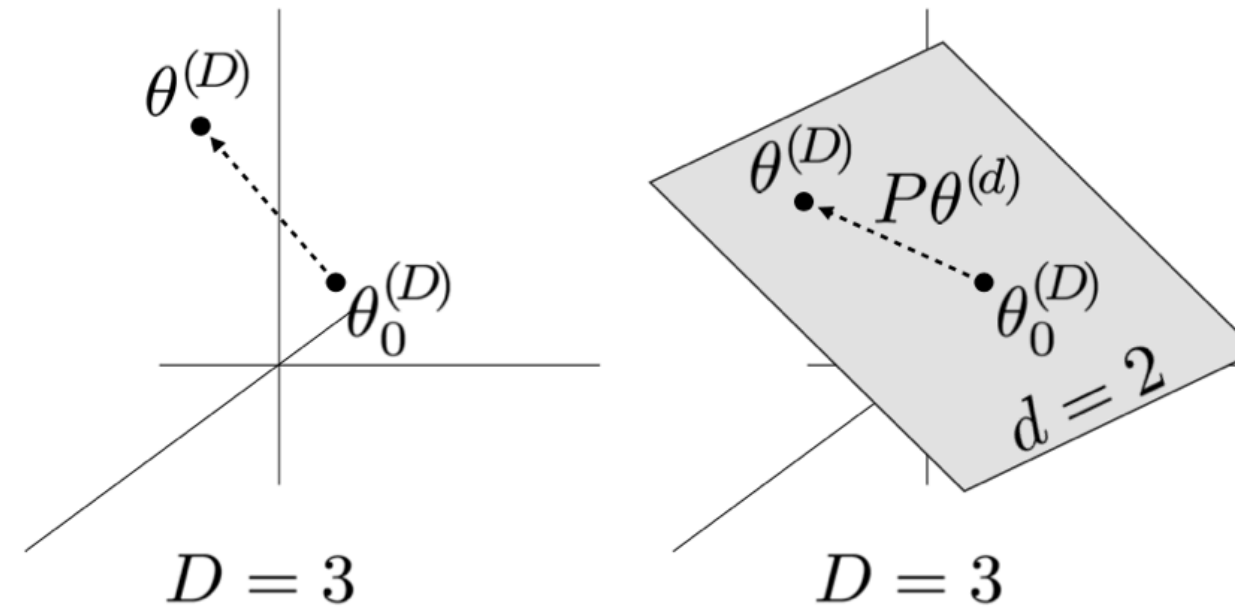
- *“I can’t quite place you as an expert of anything; what do you want to focus on next?”*

Pain point Yup. I am more of a generalist (in NNs).

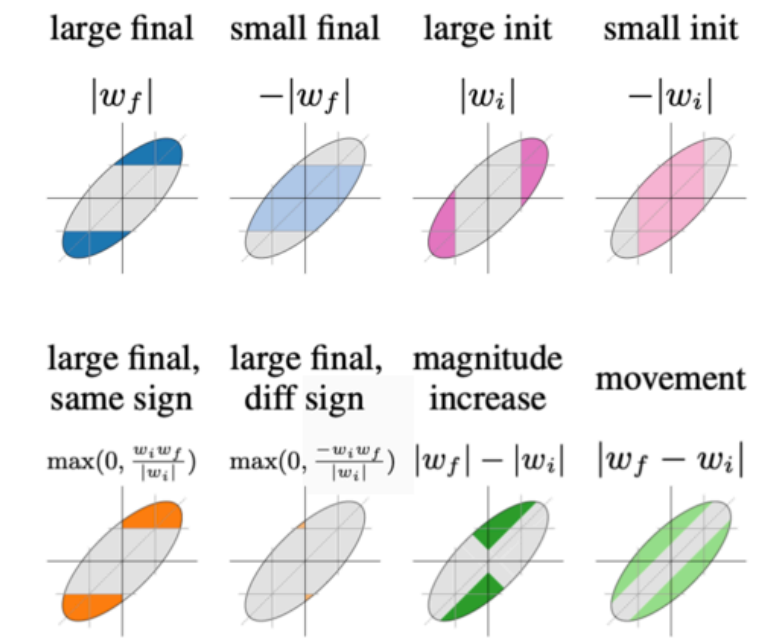
Reason



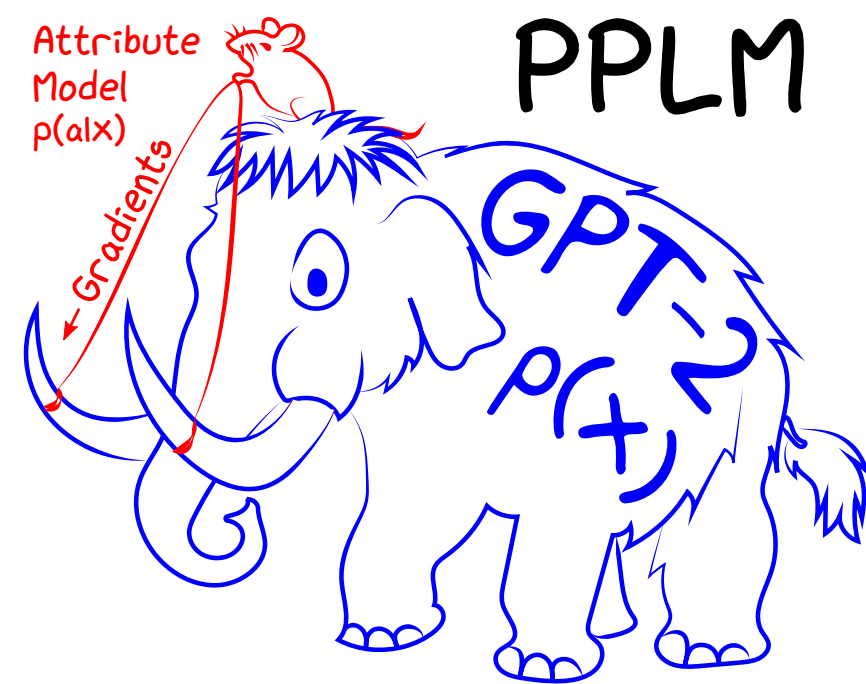
CoordConv;
NeurIPS 2018



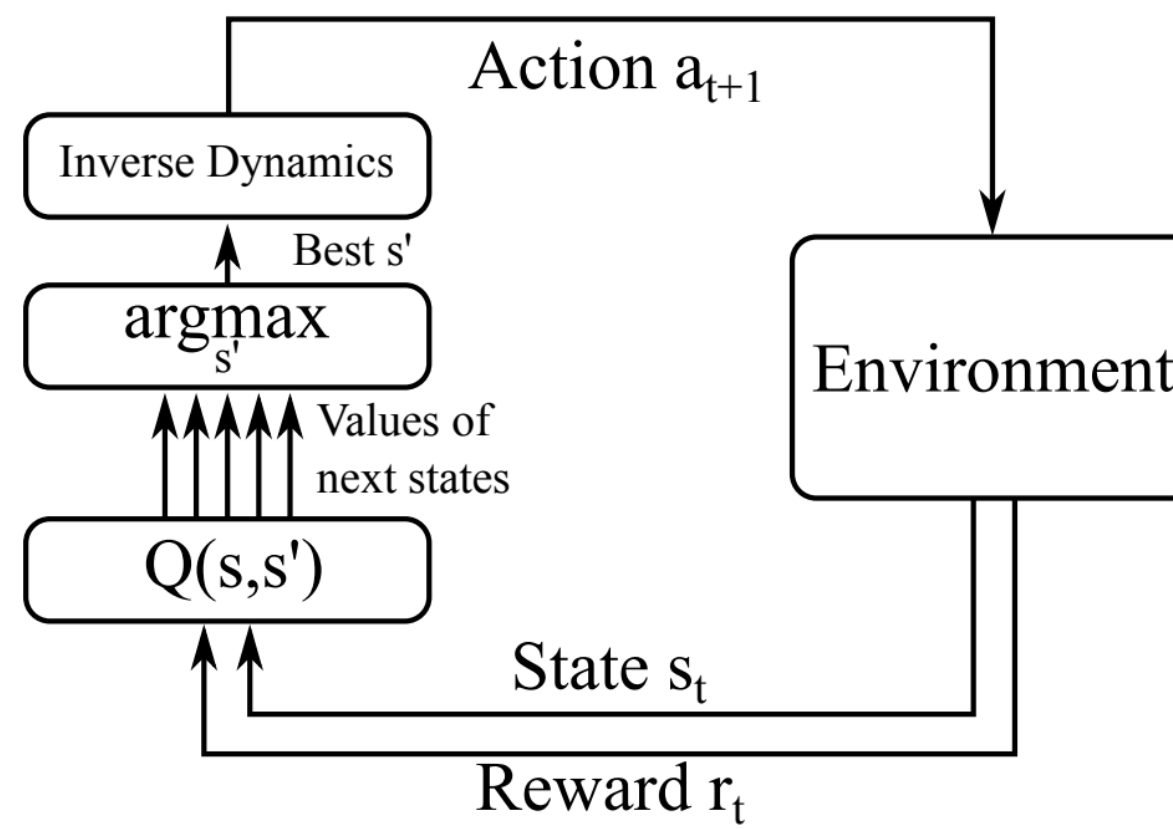
Intrinsic Dimension;
ICLR 2018



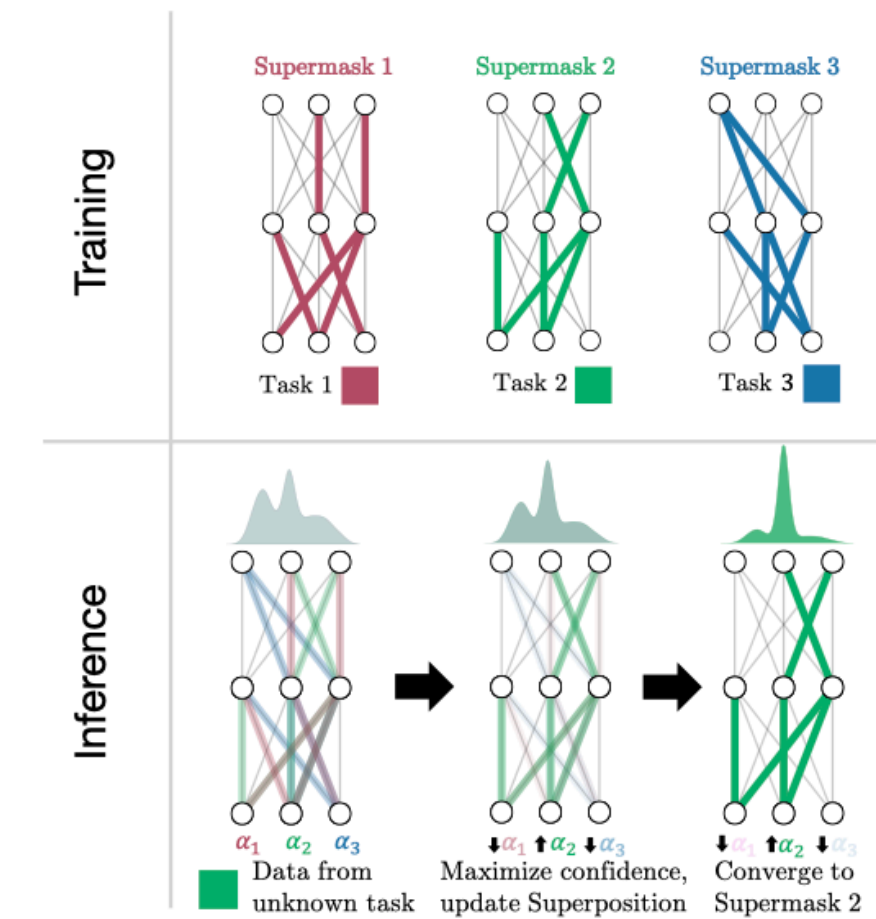
DLT;
NeurIPS 2019



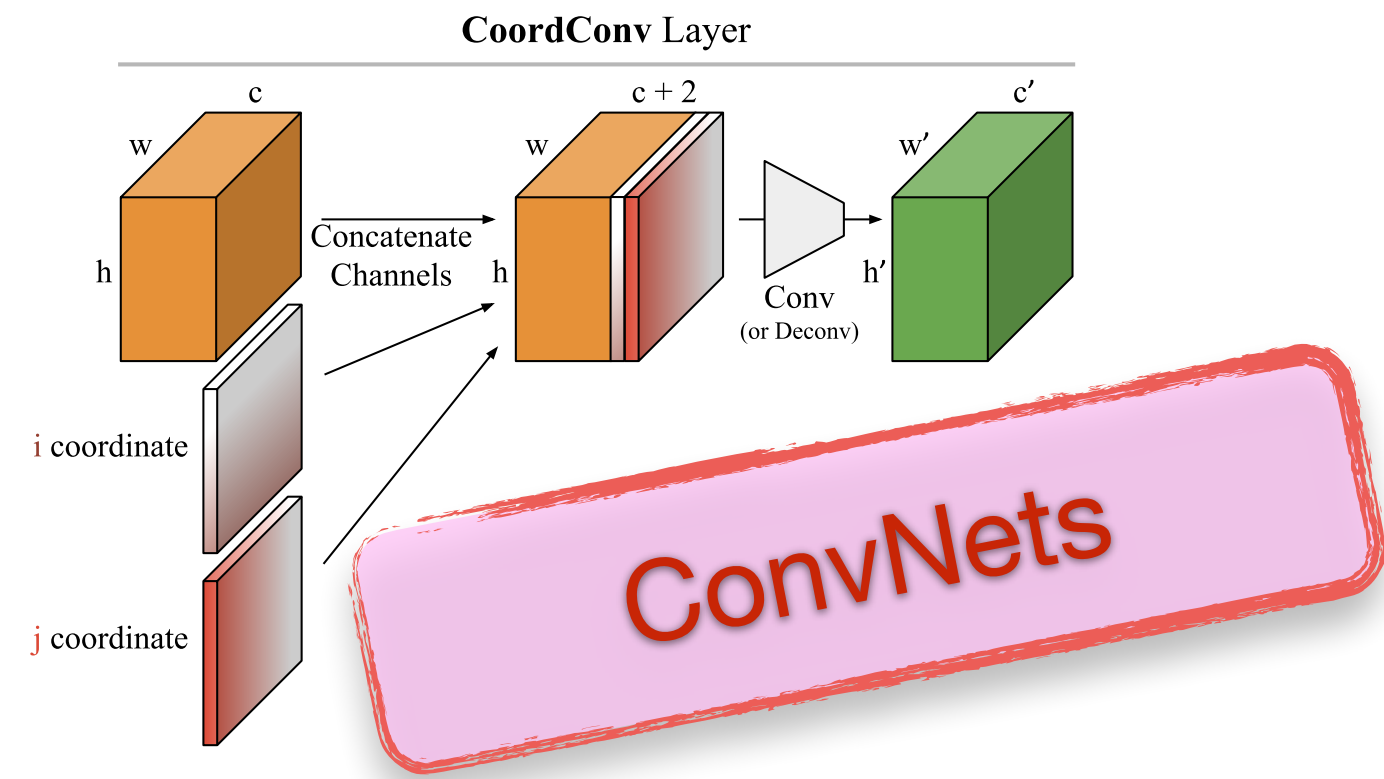
PPLM;
ICLR 2020



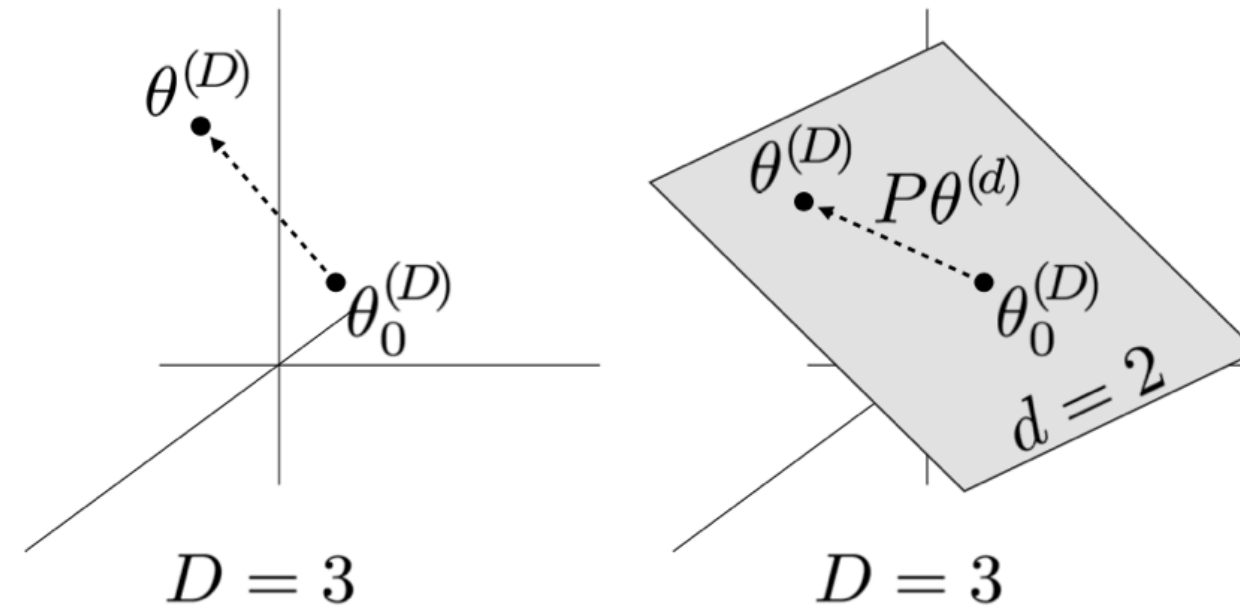
QSS;
ICML 2020



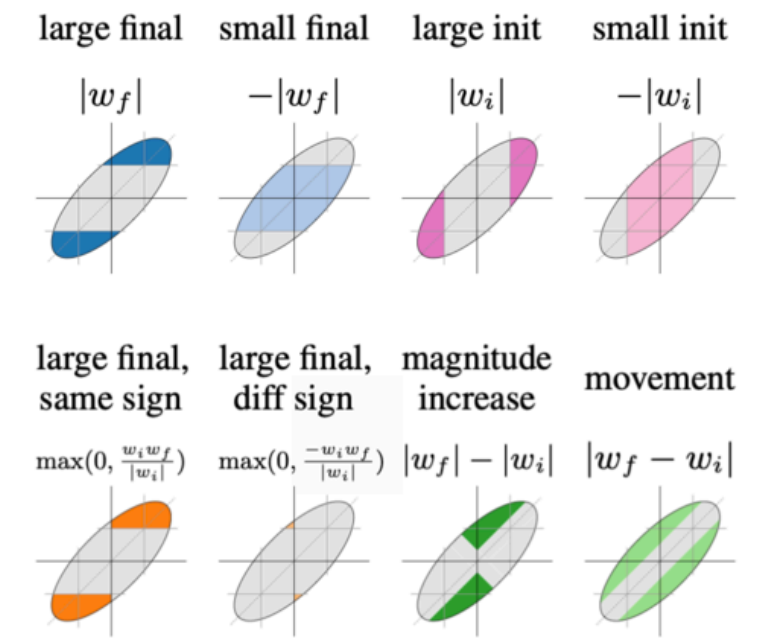
SupSup;
NeurIPS 2020



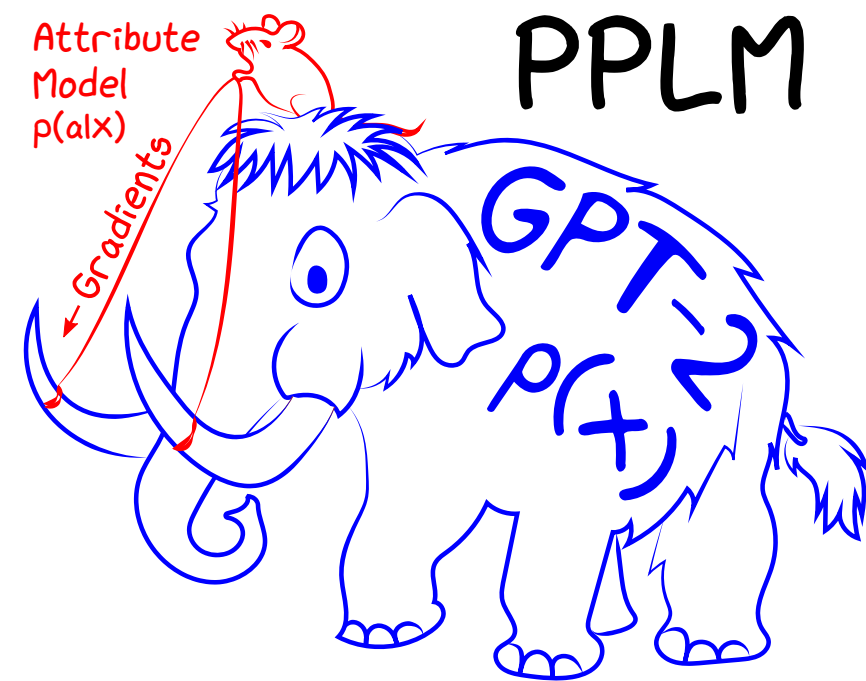
CoordConv;
NeurIPS 2018



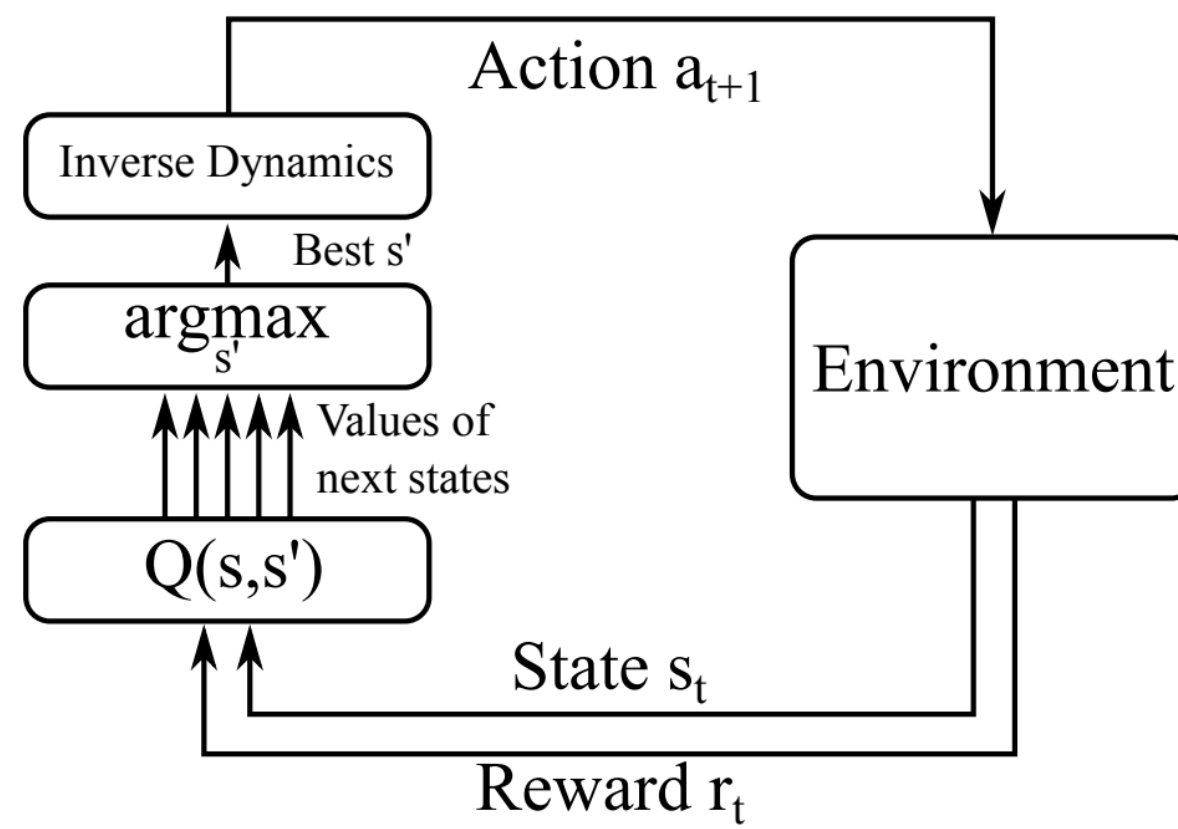
Intrinsic Dimension;
ICLR 2018



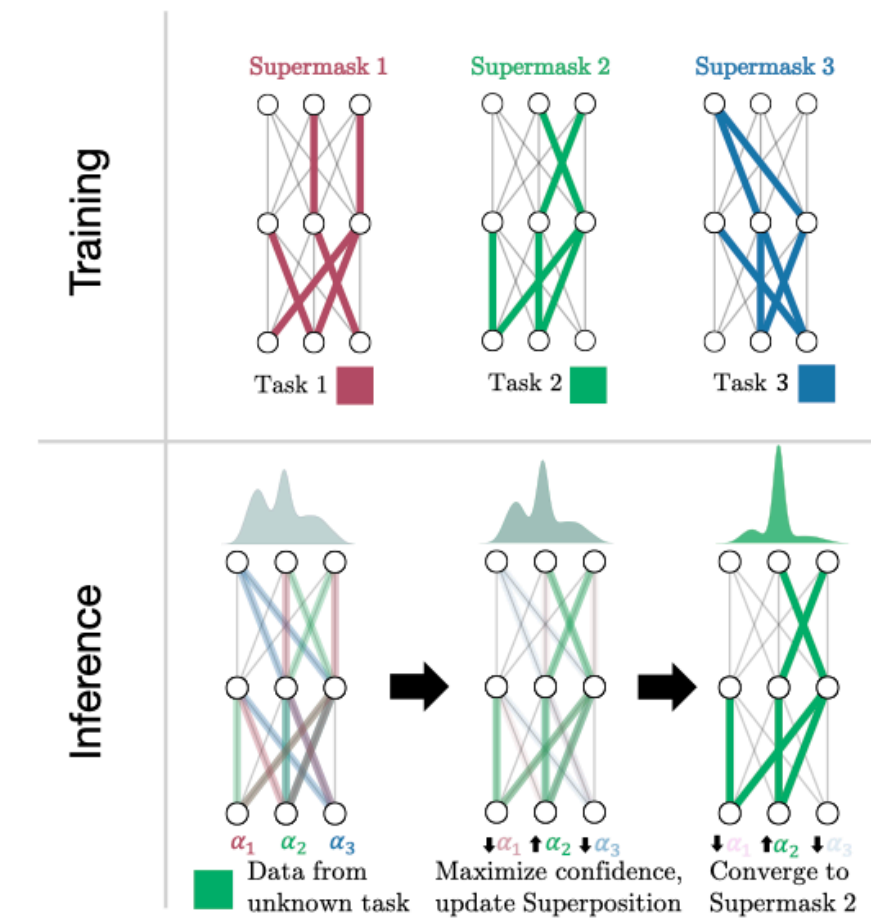
DLT;
NeurIPS 2019



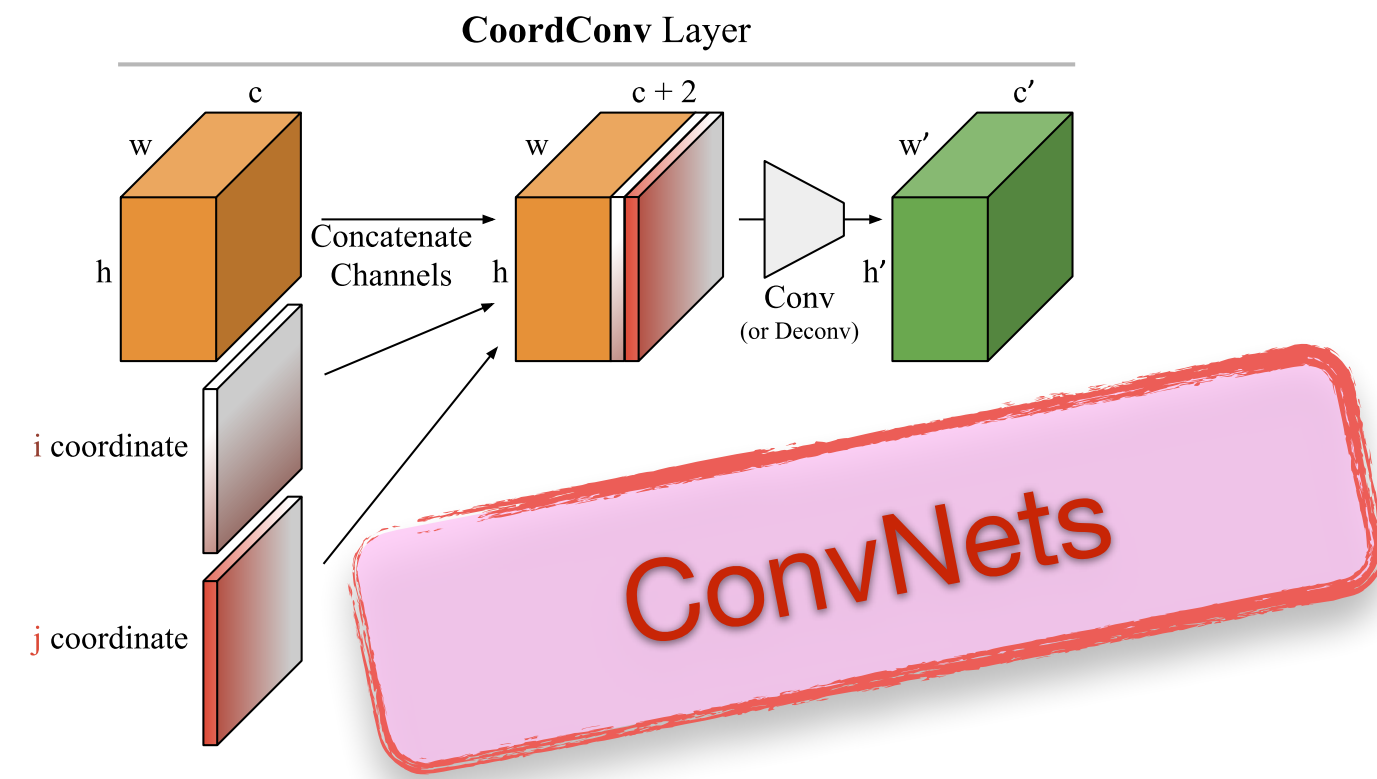
PPLM;
ICLR 2020



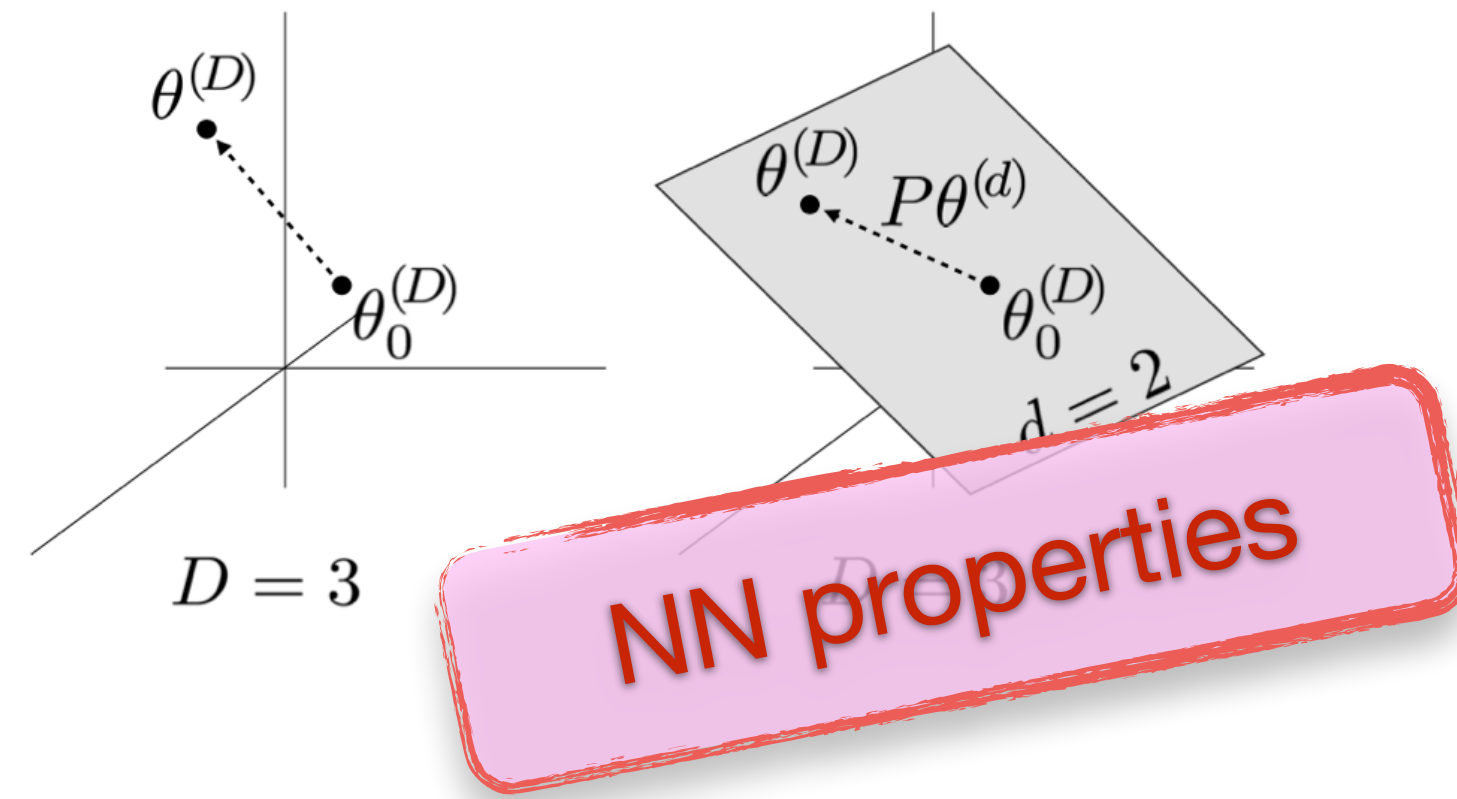
QSS;
ICML 2020



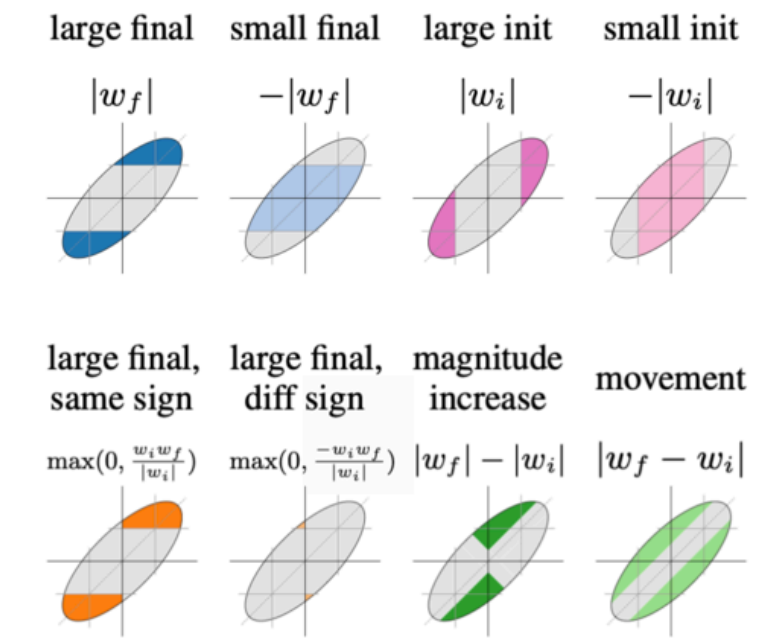
SupSup;
NeurIPS 2020



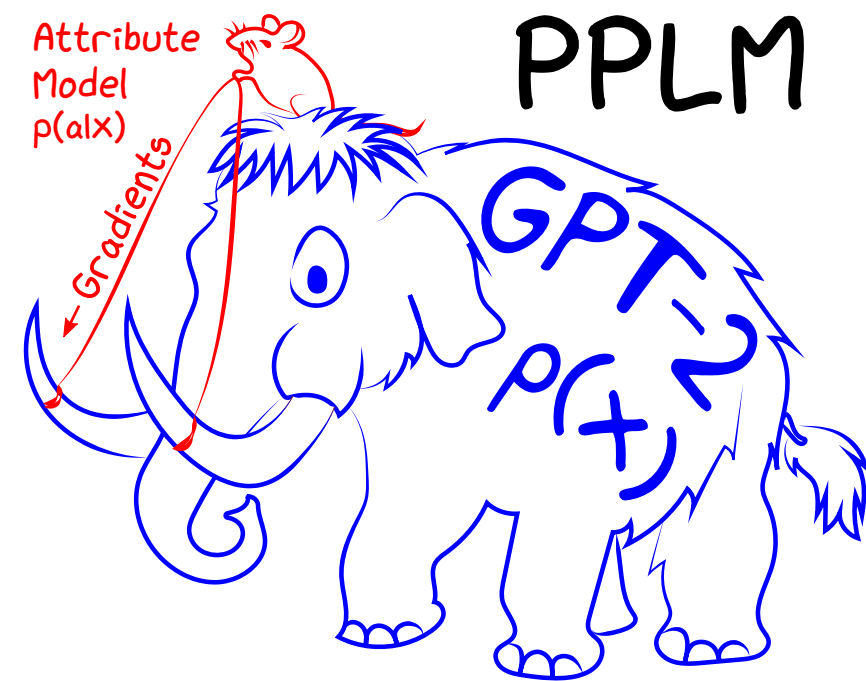
CoordConv;
NeurIPS 2018



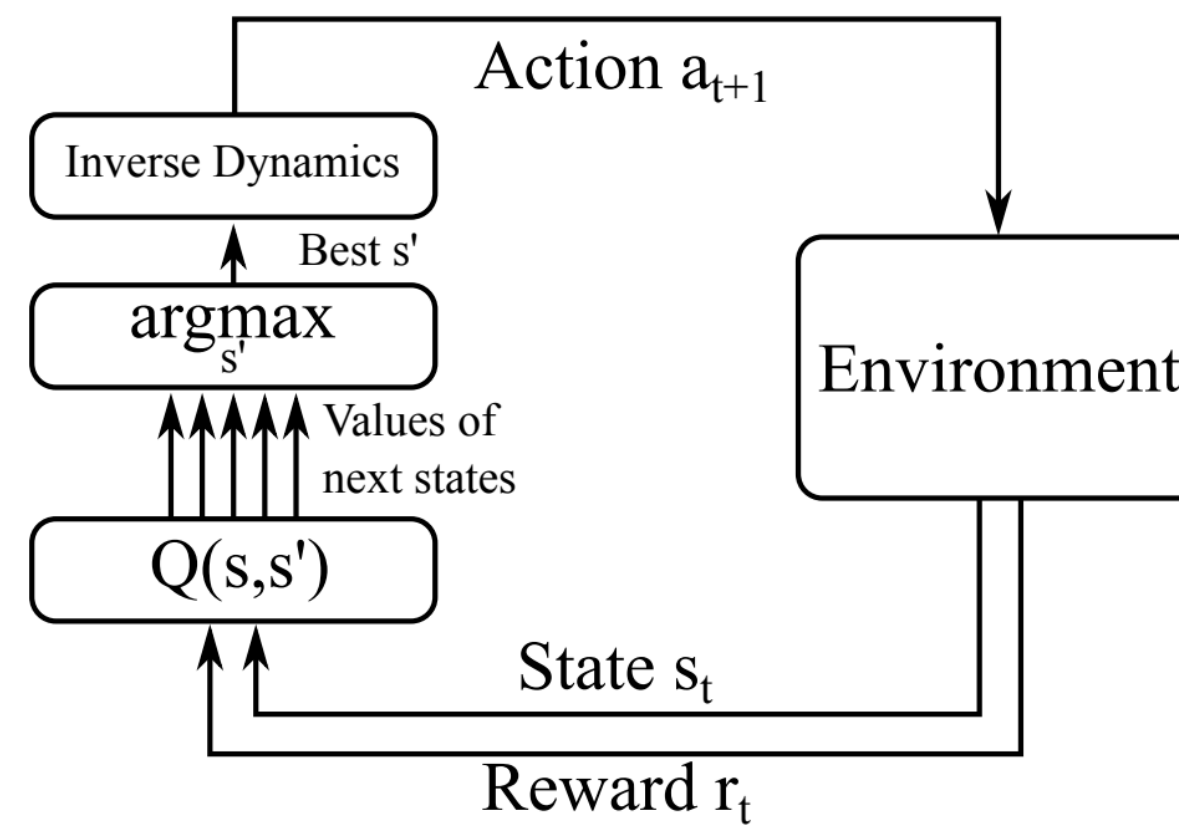
Intrinsic Dimension;
ICLR 2018



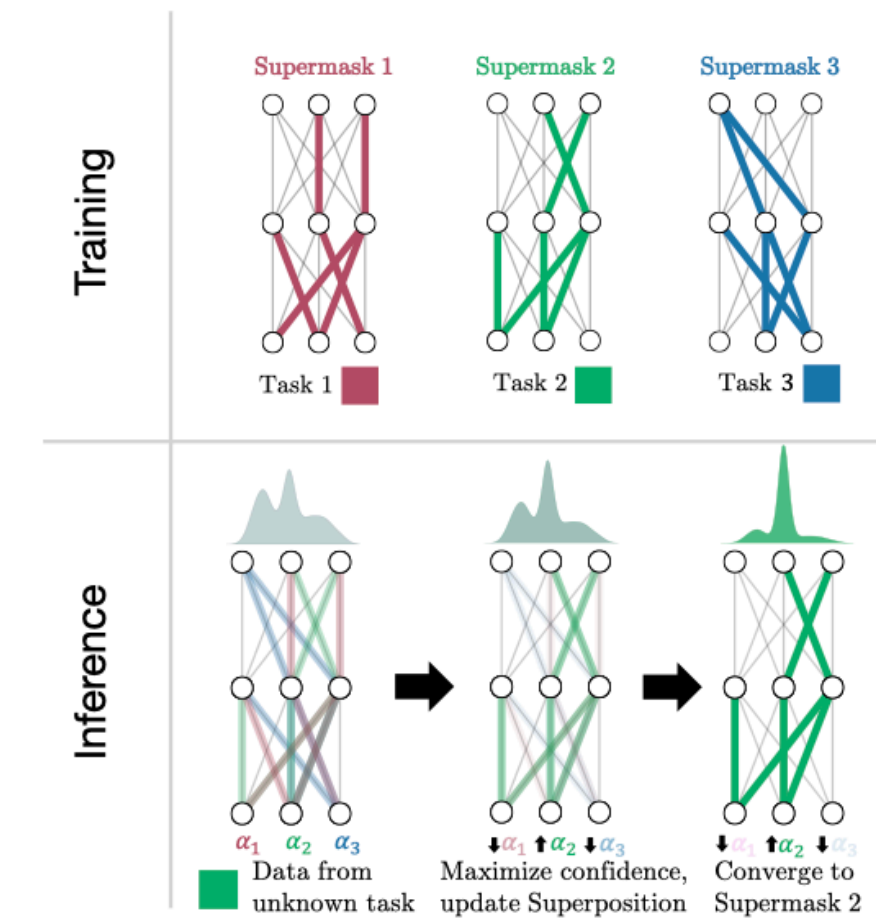
DLT;
NeurIPS 2019



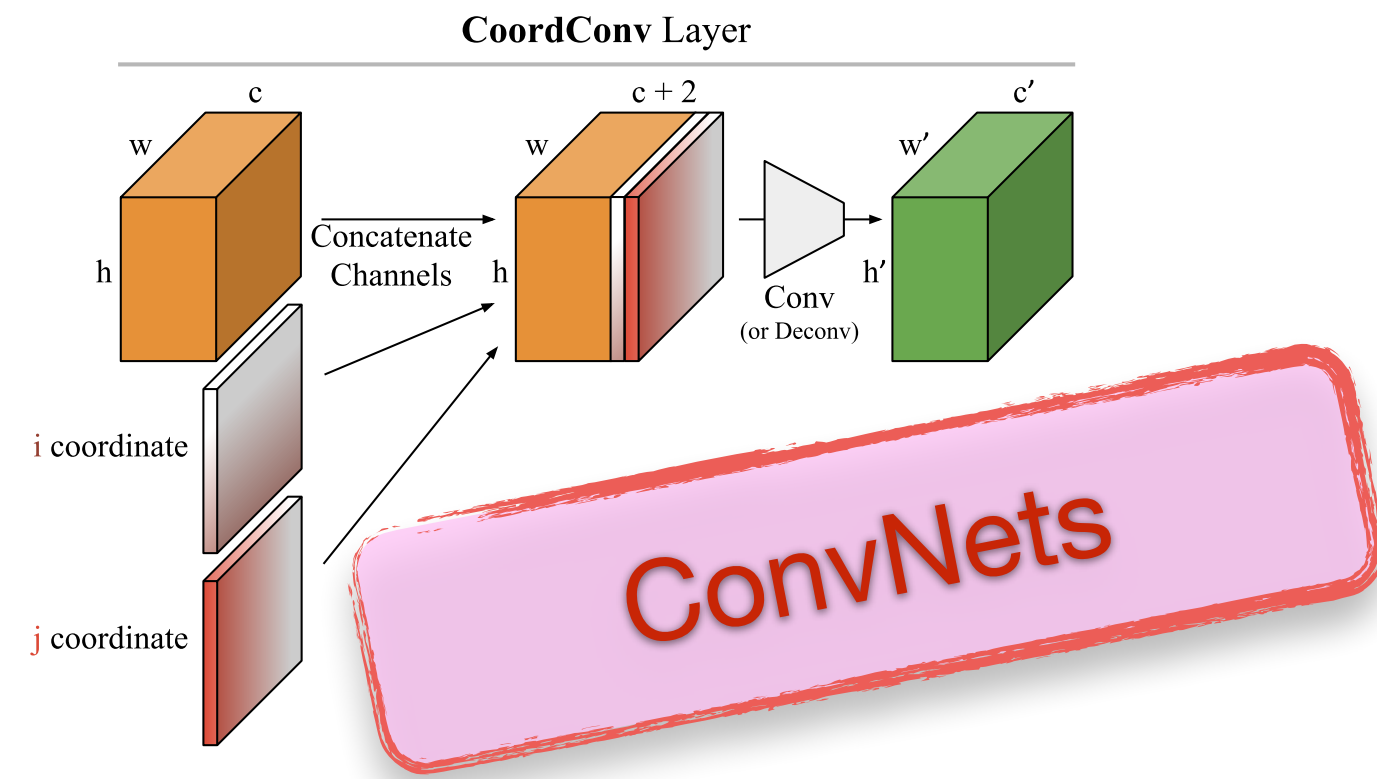
PPLM;
ICLR 2020



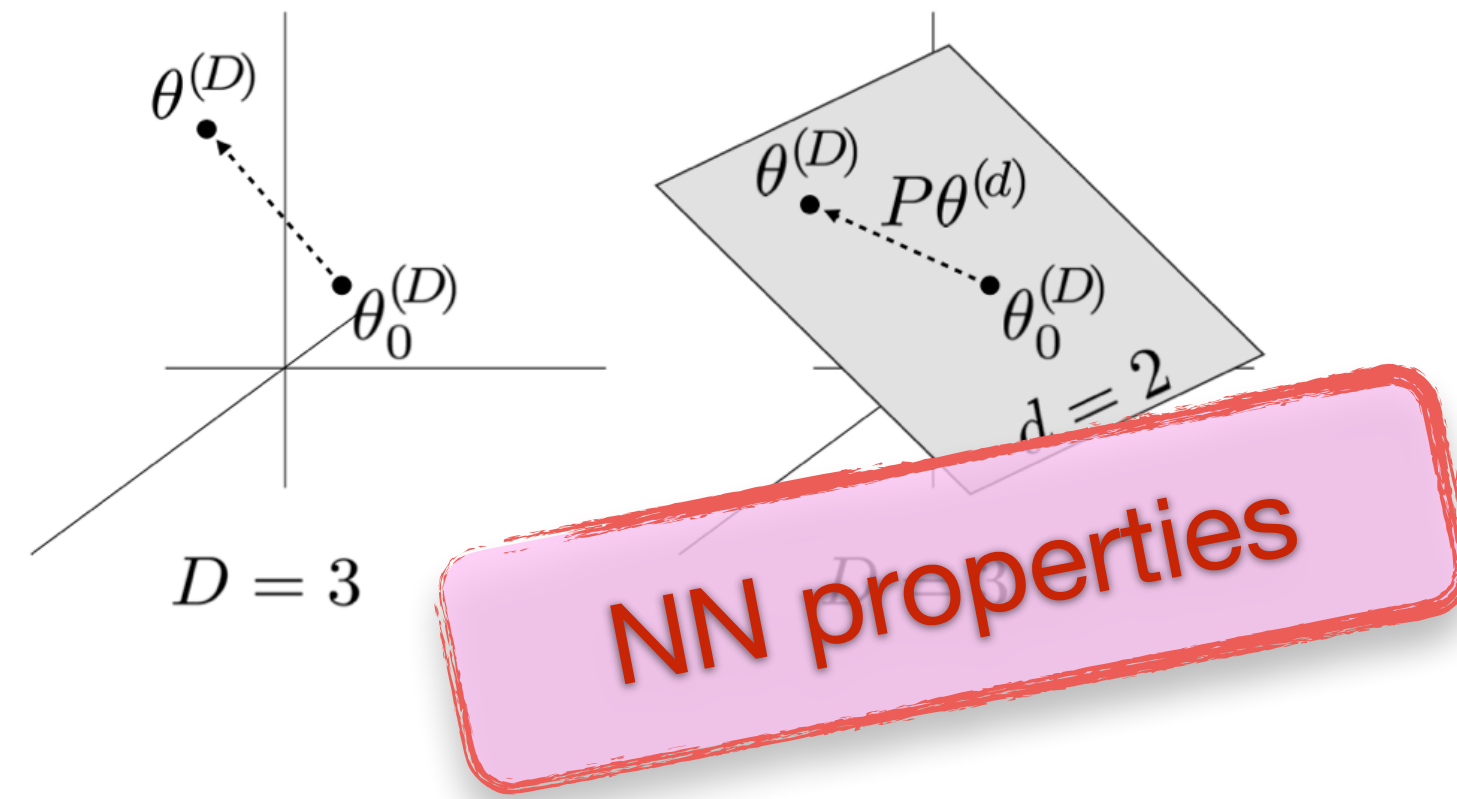
QSS;
ICML 2020



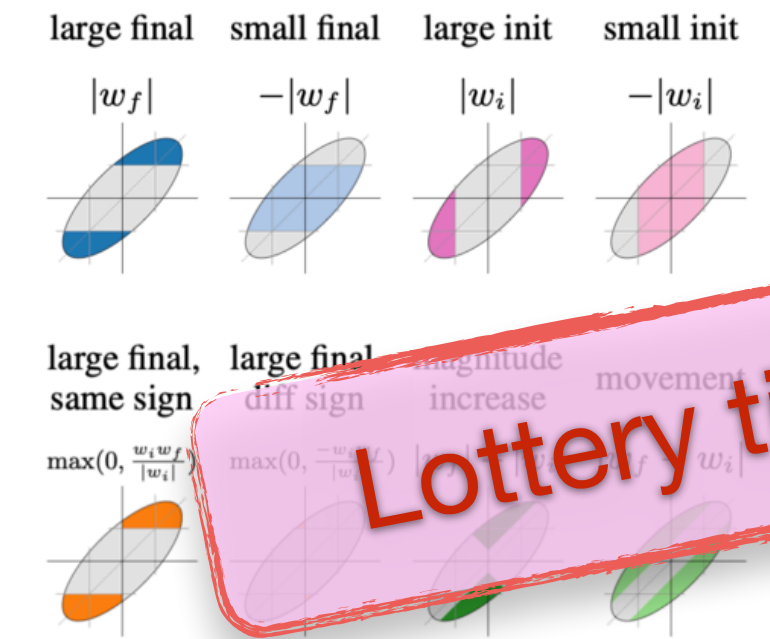
SupSup;
NeurIPS 2020



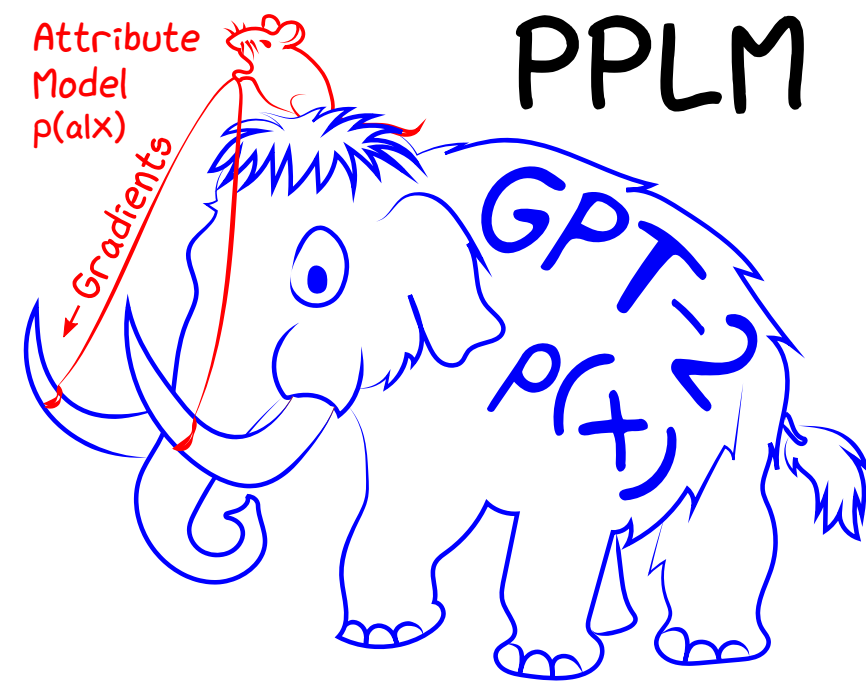
CoordConv;
NeurIPS 2018



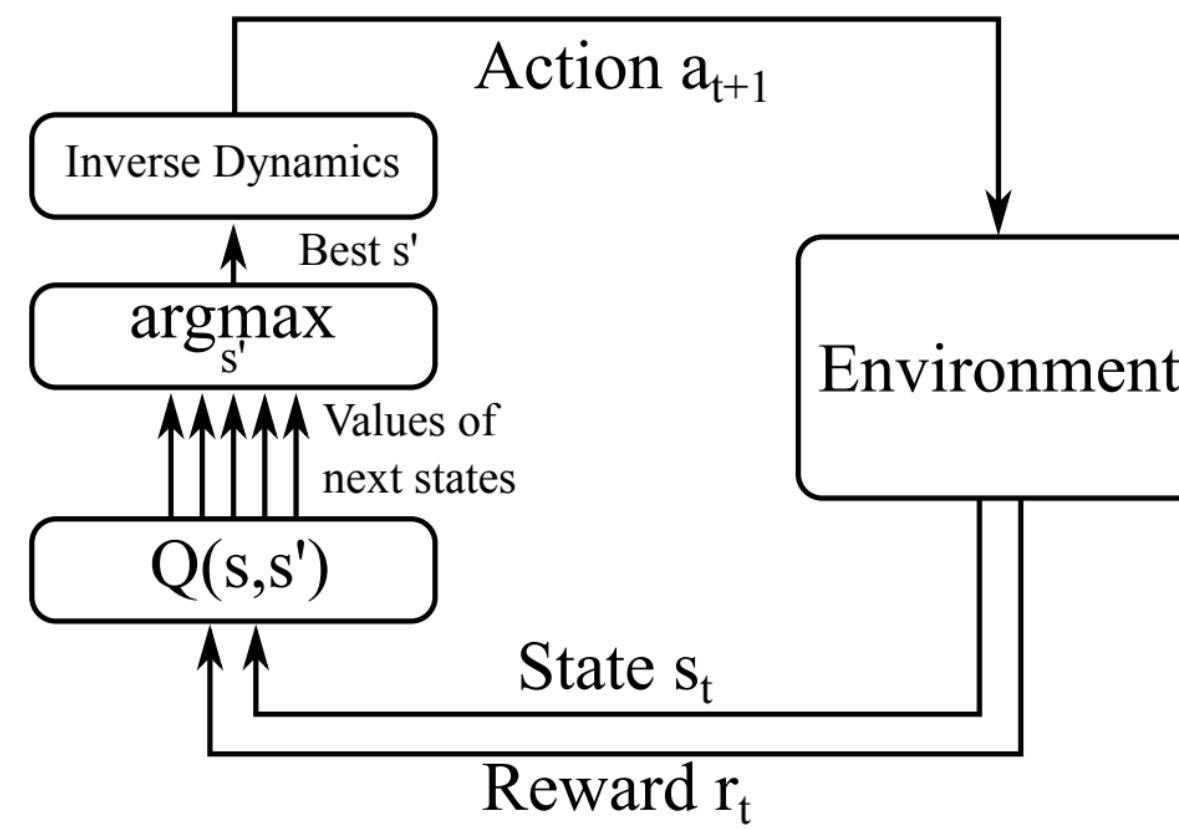
Intrinsic Dimension;
ICLR 2018



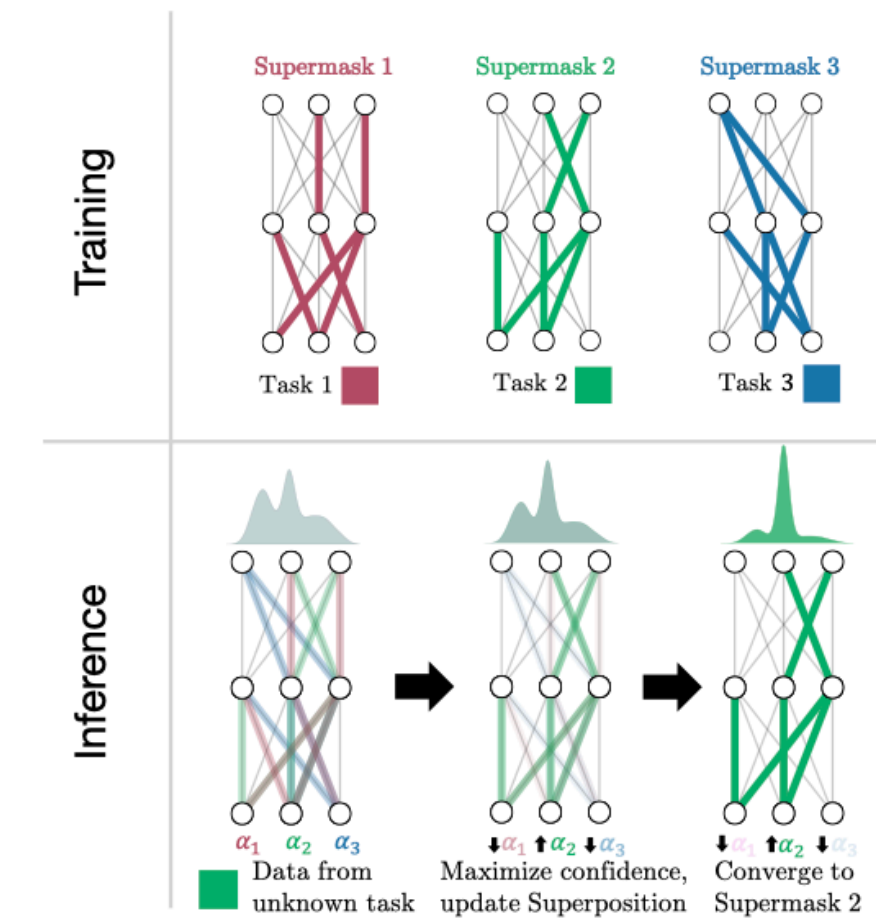
DLT;
NeurIPS 2019



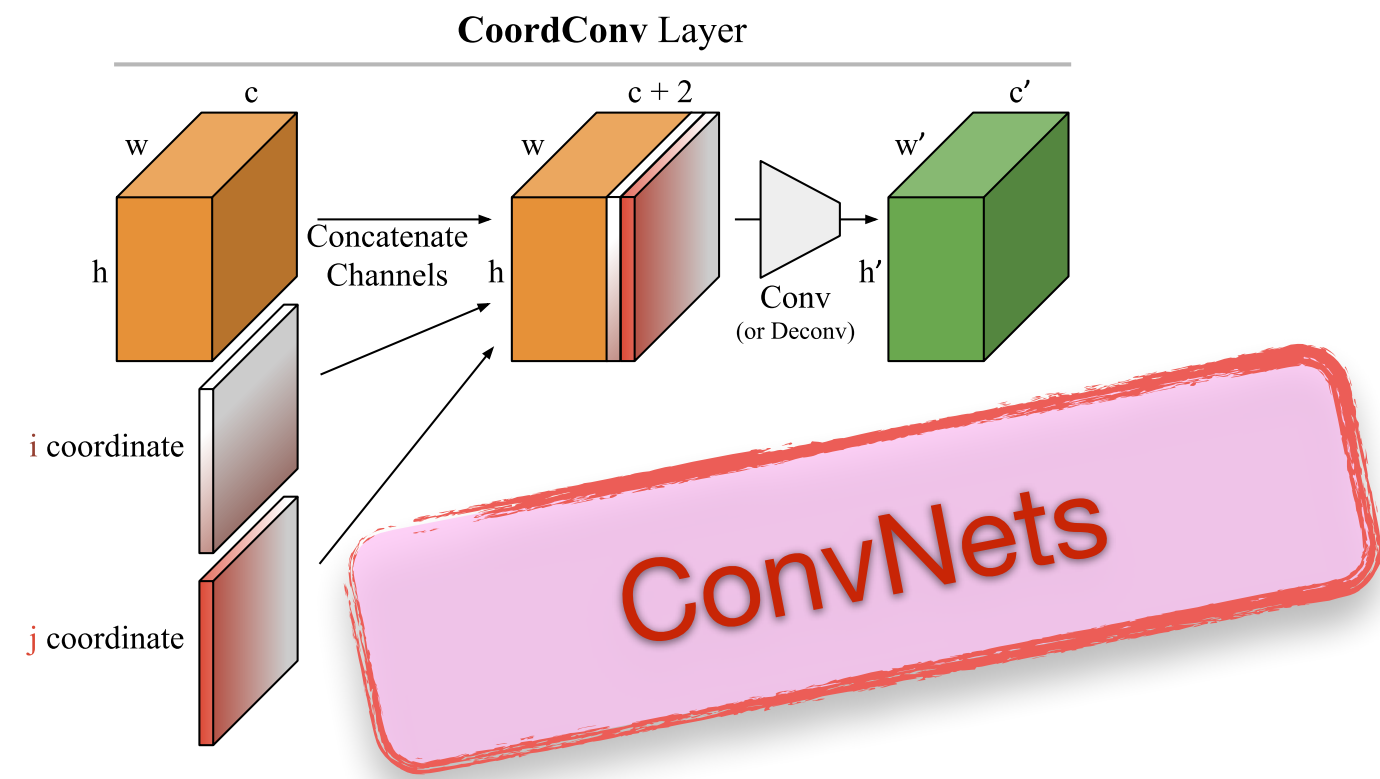
PPLM;
ICLR 2020



QSS;
ICML 2020

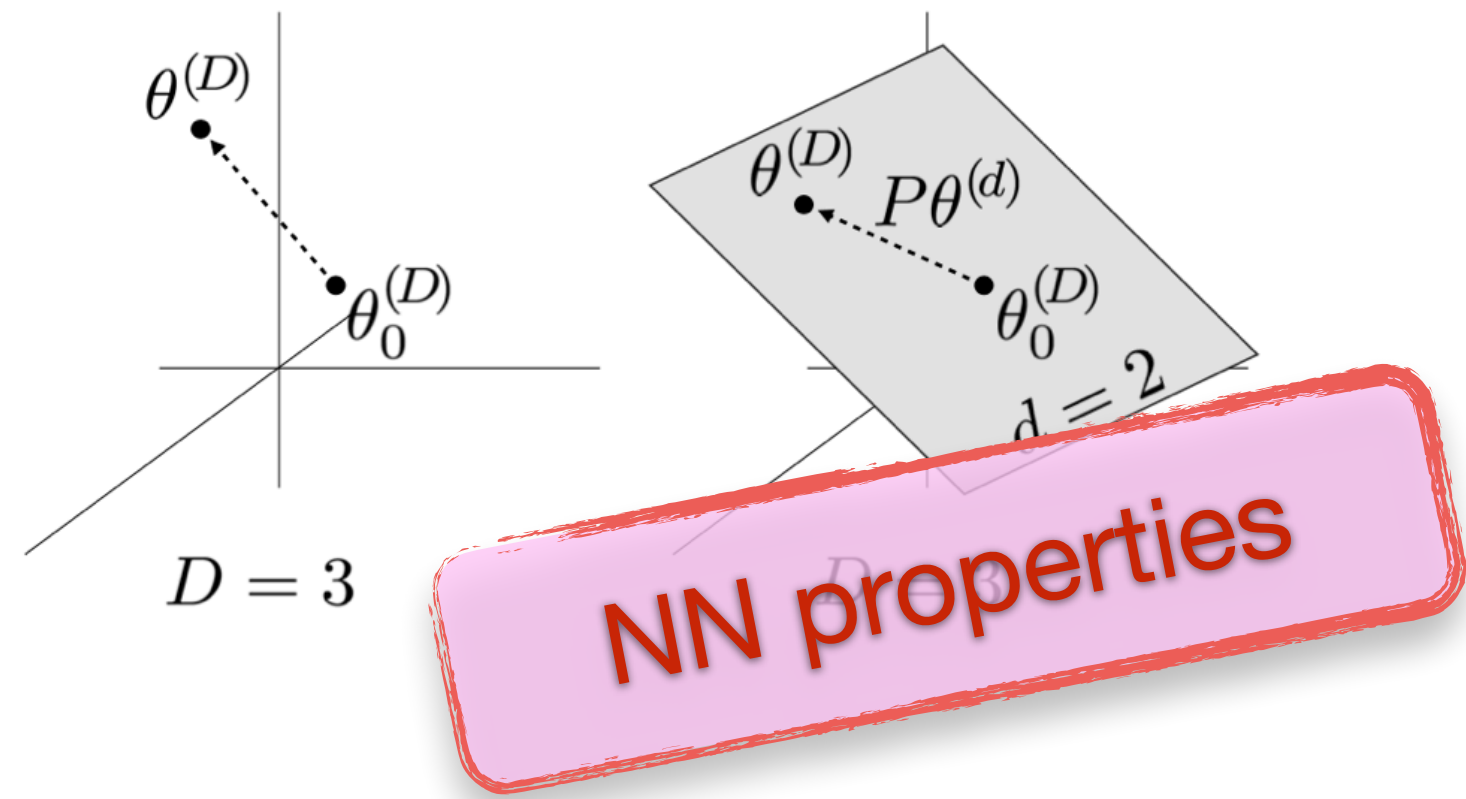


SupSup;
NeurIPS 2020



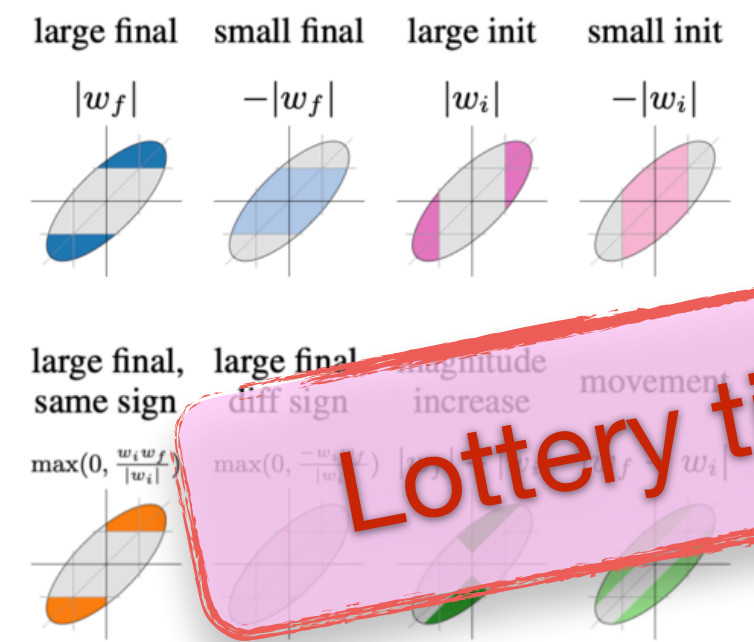
CoordConv;
NeurIPS 2018

ConvNets



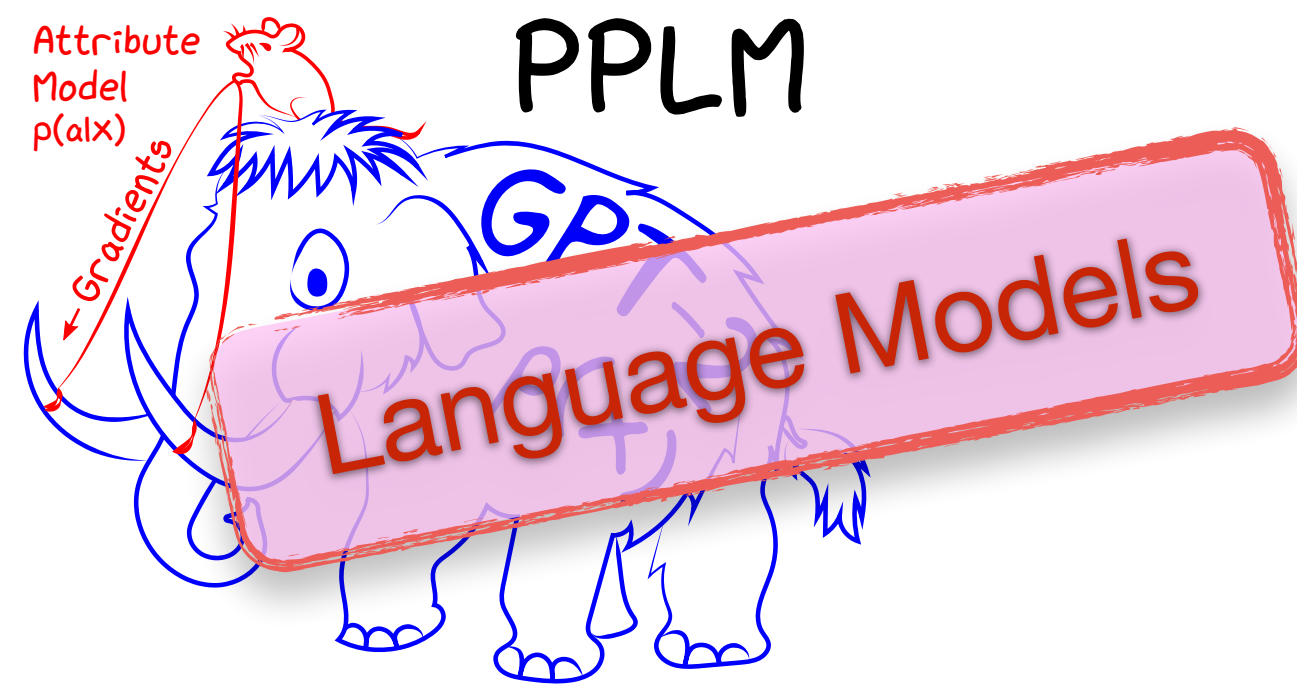
Intrinsic Dimension;
ICLR 2018

NN properties



DLT;
NeurIPS 2019

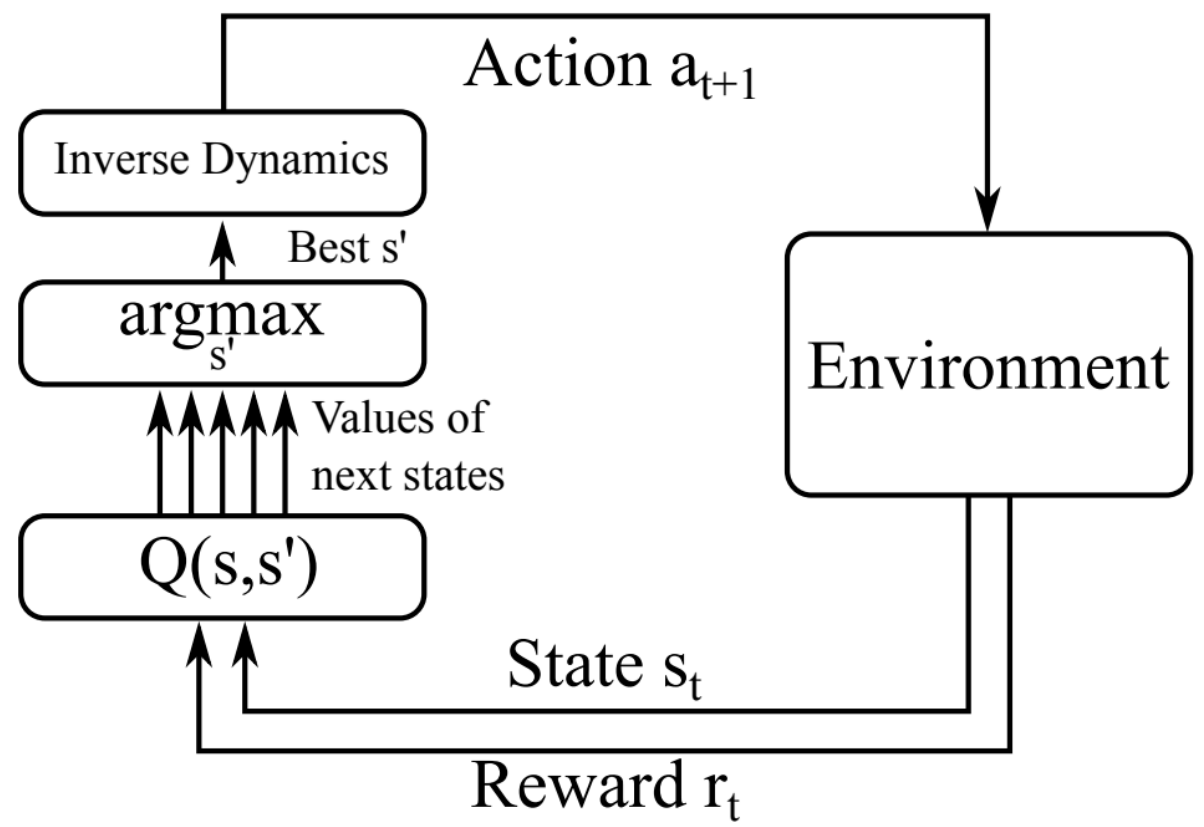
Lottery tickets



PPLM;
ICLR 2020

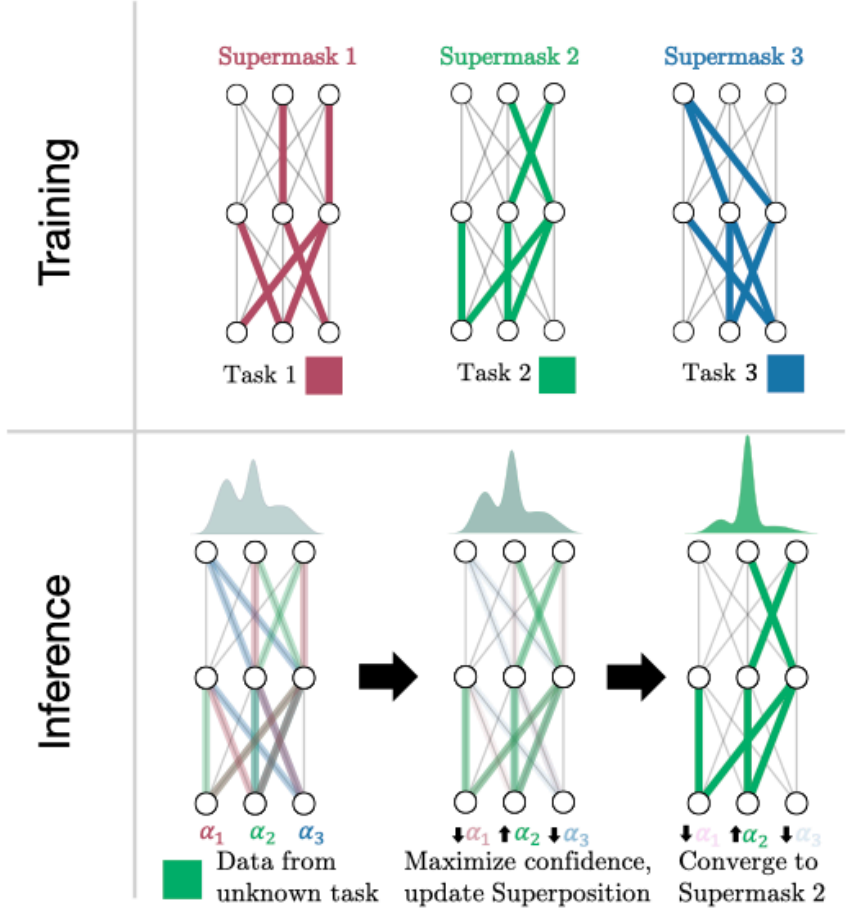
PPLM

Language Models



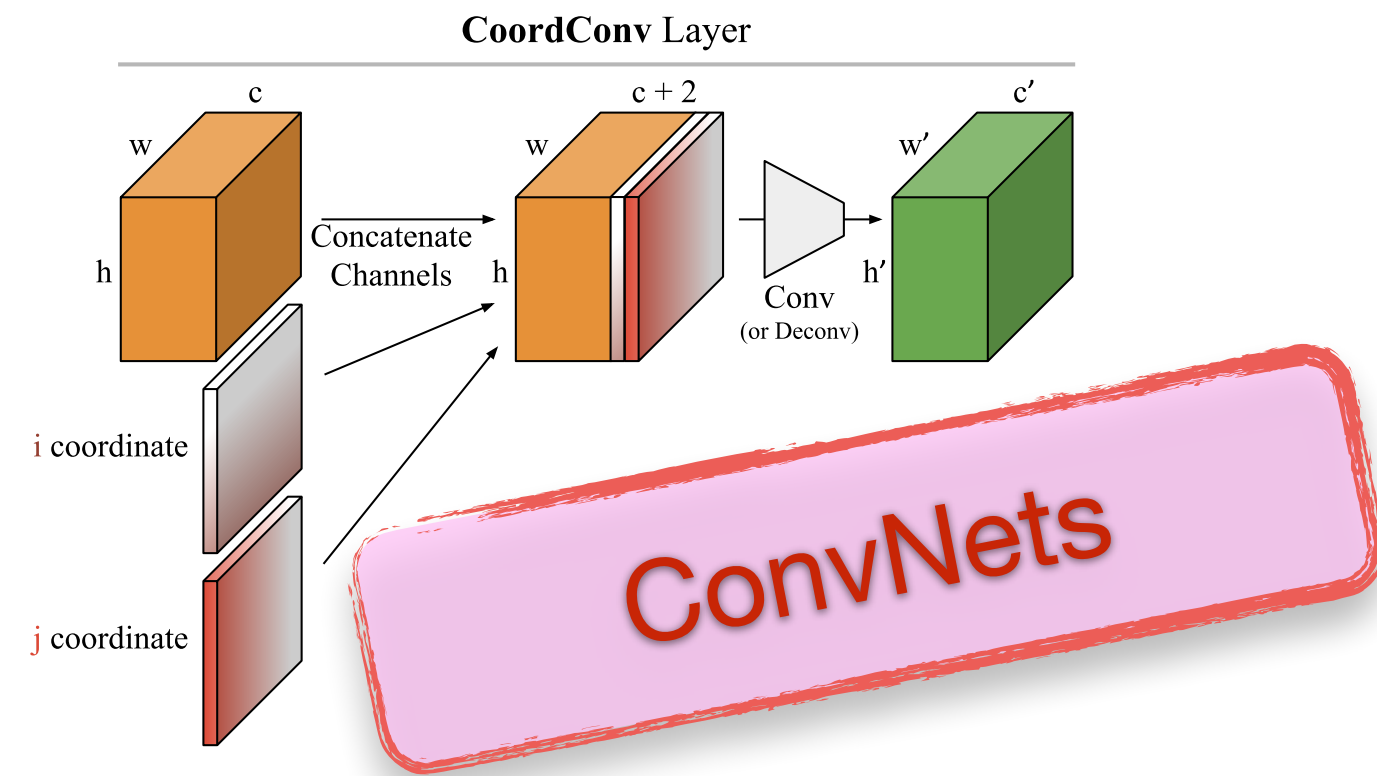
QSS;
ICML 2020

QSS

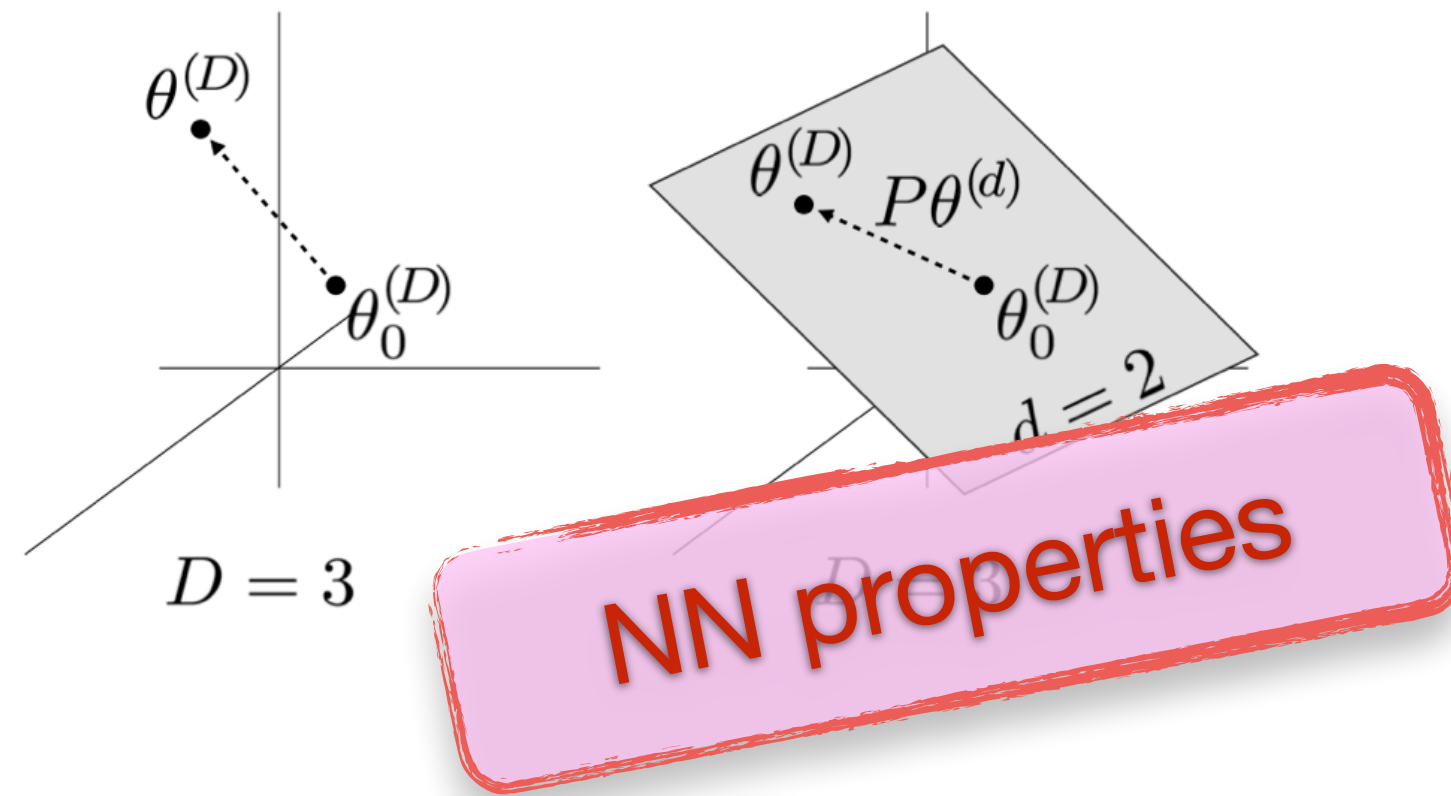


SupSup;
NeurIPS 2020

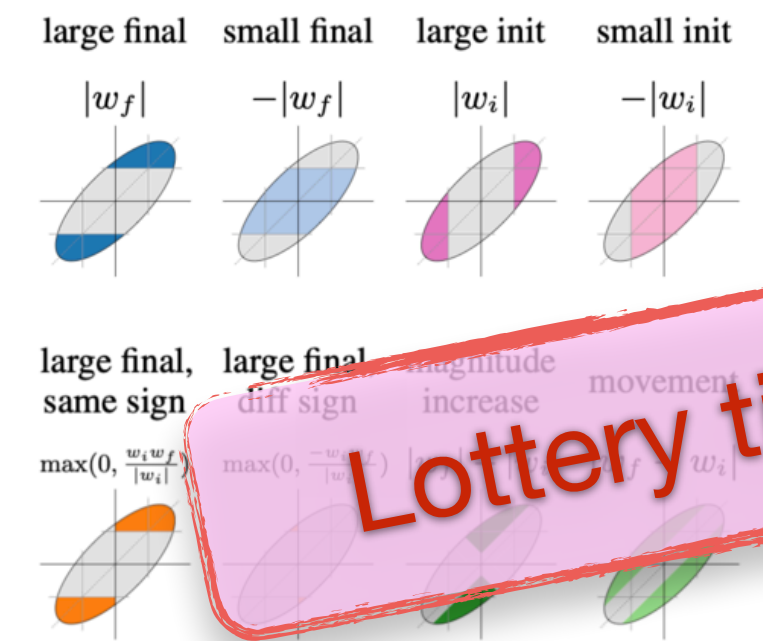
SupSup



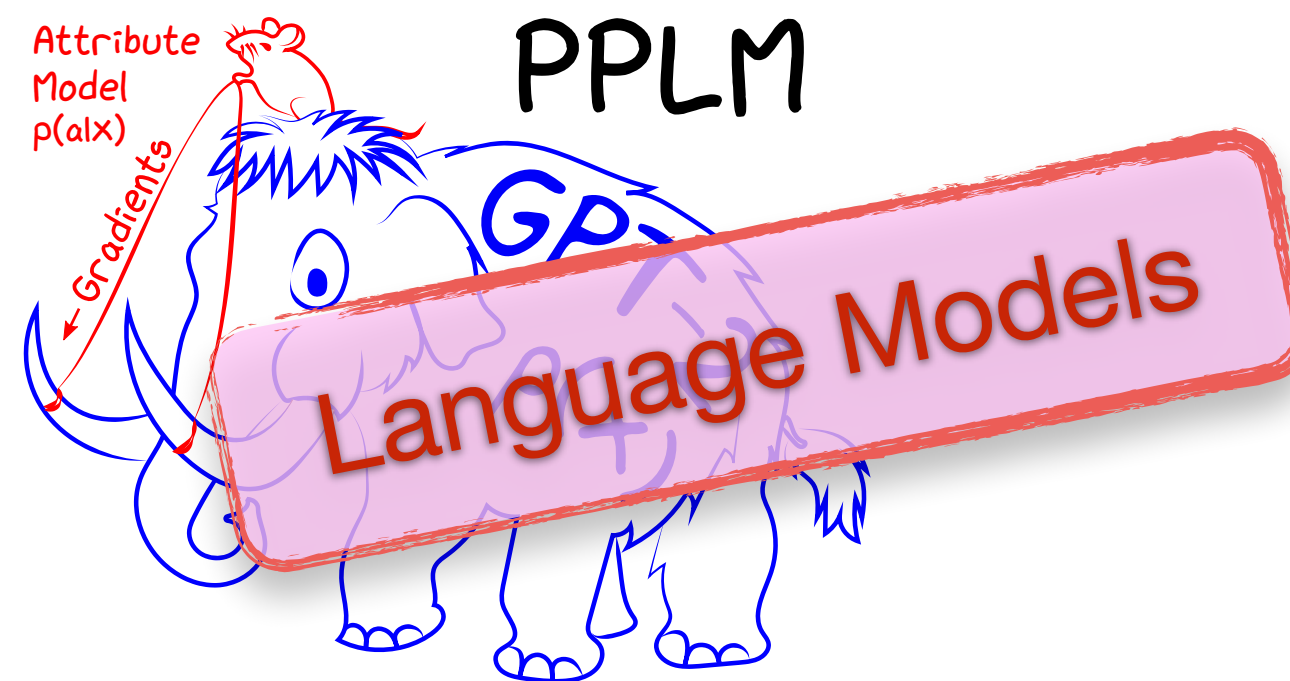
CoordConv;
NeurIPS 2018



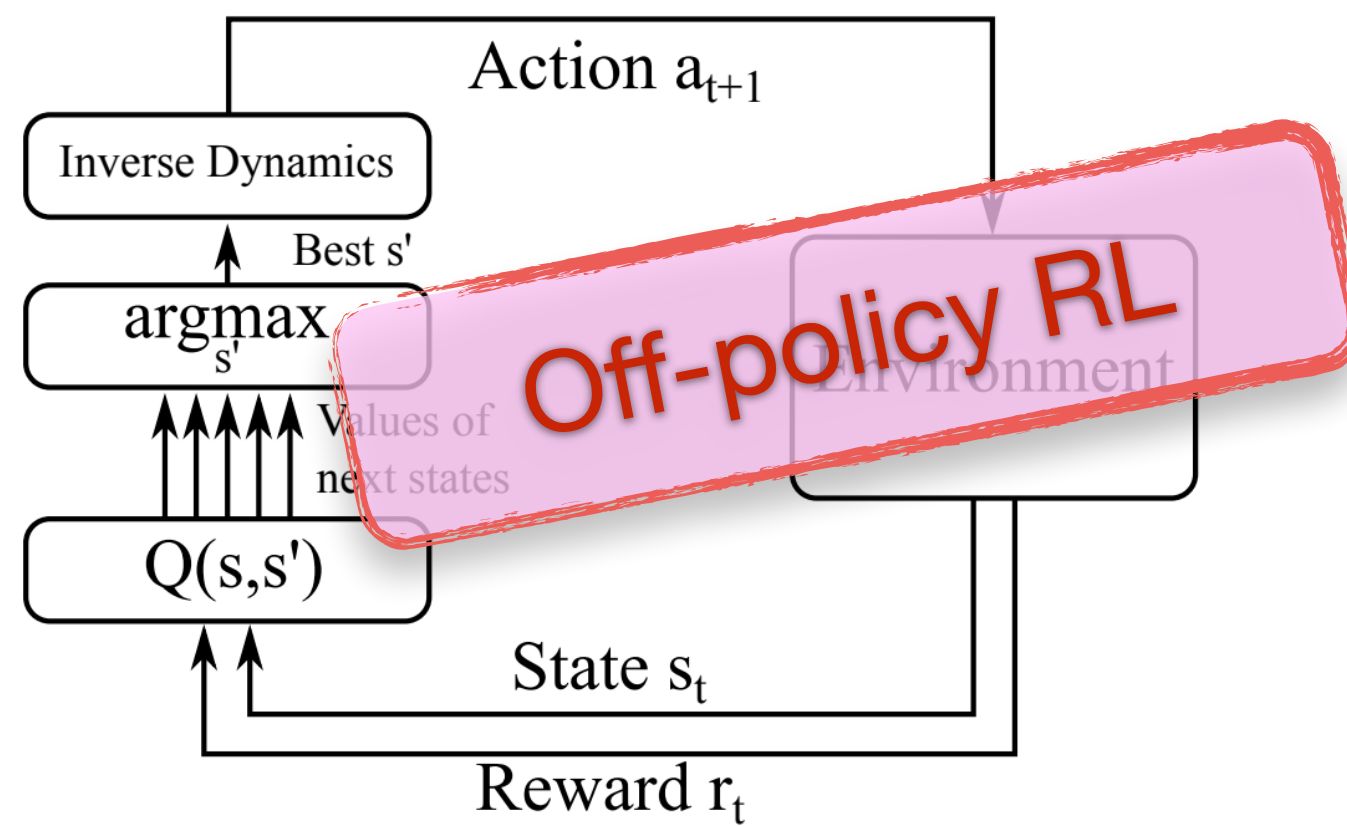
Intrinsic Dimension;
ICLR 2018



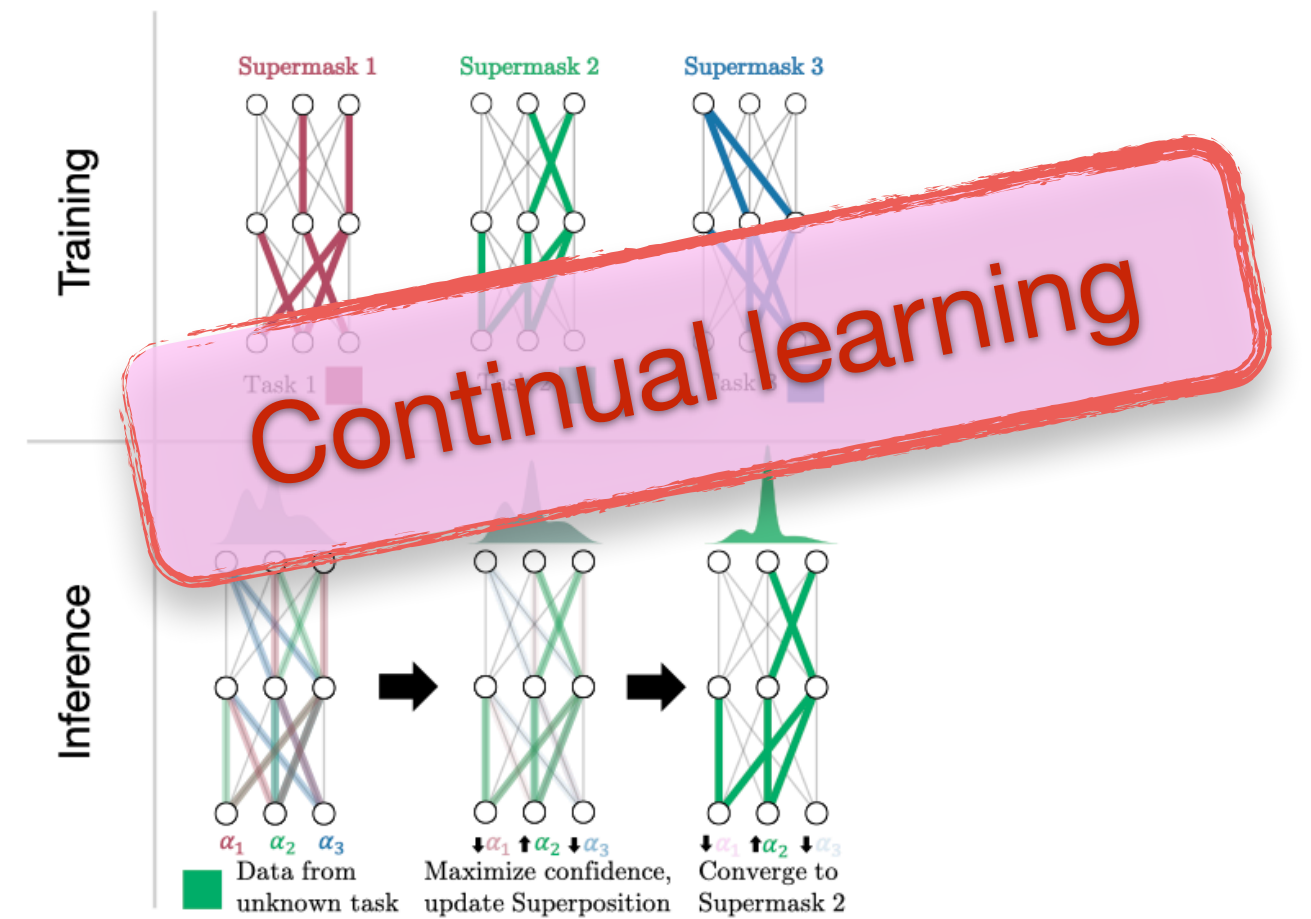
DLT;
NeurIPS 2019



PPLM;
ICLR 2020



QSS;
ICML 2020



SupSup;
NeurIPS 2020

- *“I can’t quite place you as an expert of anything; what do you want to focus on next?”*

Pain point Yup. I am more of a generalist (in NNs).

Reason

- *“I can’t quite place you as an expert of anything; what do you want to focus on next?”*

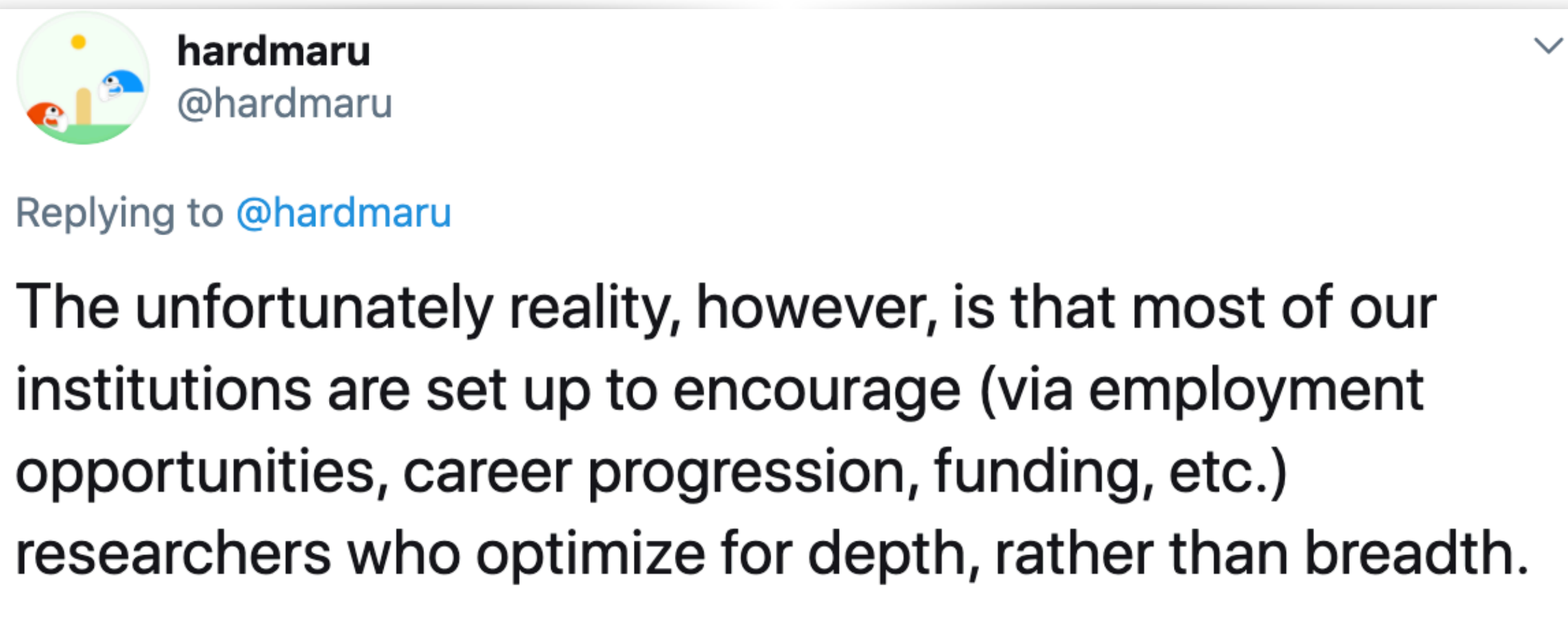
Pain point Yup. I am more of a generalist (in NNs).

Reason The hyperspecialized societal culture has seeped into research.

- *“I can’t quite place you as an expert of anything; what do you want to focus on next?”*

Pain point Yup. I am more of a generalist (in NNs).

Reason The hyperspecialized societal culture has seeped into research.



<https://twitter.com/hardmaru/status/1309313448262512640?s=20>

Apparently there are two paths for AI research

- Academia, often in one place
- ~~Industry, likely on cycles~~ Flawed

Wait what about academia?



Charles Isbell @isbellHFh · Sep 10

2/n

The secret to understating how to get into a PhD program is understanding how faculty think: we all believe we can tell within 15 seconds whether you're good enough to be one of us. Our evaluations of possible students depends basically on approximating knowing you:



2



3



36



Charles Isbell

@isbellHFh

Replying to @isbellHFh

postscript:

"But why?" you ask?

Simple. The entire system is designed to minimize false positives. Who cares about false negatives when you have 5-50 times more applicants than slots?

...and, yes, it is FAR FAR FAR worse for faculty positions. I've got numbers and everything.

11:05 AM · Sep 10, 2020 · Twitter Web App

<https://twitter.com/isbellHFh/status/1304110091873013761?s=20>

What now?

What now?

Change yourself (*easy*), or change the system (*hard*).

- *“You were the middle author in a lot of papers, what exactly were your contributions?”*

Pain point

The increasingly messy credit assignment problem.

Reason

There's now too much (quantifiable) gains associated with ML papers.

- *“You were the middle author in a lot of papers, what exactly were your contributions?”*

Pain point

The increasingly messy credit assignment problem.

Reason

There's now too much (quantifiable) gains associated with ML papers.

Fix (direct)

Break the pattern! Work on a few first-author papers.

- *“Almost all of your papers have Jason on it, who are you removed from him?”*

Pain point

Behind any mildly successful woman there is a white man. (No sarcasm. It's true!)

Reason

Stereotype?

Fix (direct)

Break the pattern! Work on a few papers with someone else, or by myself.

- *“I can’t quite place you as an expert of anything; what do you want to focus on next?”*

Pain point Yup. I am more of a generalist (in NNs).

Reason The hyperspecialized societal culture has seeped into research.

Fix (direct) Break the pattern! Focus, and make a name in one small thing first.

What now?

~~Change yourself (easy),~~ or change the system (*hard*).

What now?

~~Change yourself (easy),~~ or change the system (*hard*).

This grumpy talk is not to address the outcome of a failed job search.

What now?

 Change yourself (easy), or change the system (*hard*).

This grumpy talk is not to address the outcome of a failed job search.

I am really speaking of this collective misery we are all having here.

Even for those who have a job.

What now?

 Change yourself (easy), or change the system (*hard*).

This grumpy talk is not to address the outcome of a failed job search.

I am really speaking of this collective misery we are all having here.

Even for those who have a job.

I am sure you are facing similar pain points in your, e.g. perf reviews—that there are parts of you that **do not fit the rubric**. What do you do? Do you change yourself? Do you try to change the system? It is really a universal question.

Pain point #2: I still aspire to do science somewhere. But wait, the kind of science that used to inspire me has changed.

Even when I had a RS job I felt there were a lot of misconceptions.

Misconception #1: Talent vs. Opportunity

Misconception #2: Individualistic vs. Collective

Misconception #3: Competitor vs. Collaborator

Misconception #1: Talent vs. Opportunity

Misconception #2: Individualistic vs. Collective

Misconception #3: Competitor vs. Collaborator



Misconception #1: Talent vs. Opportunity

Misconception #2: Individualistic vs. Collective

Misconception #3: Competitor vs. Collaborator



Misconception #1: Talent vs. Opportunity

Misconception #2: Individualistic vs. Collective

Misconception #3: Competitor vs. Collaborator



ML Collective

~~Change yourself (easy),~~ or change the system (*hard*).

- *“Almost all of your papers have Jason on it, who are you removed from him?”*

Pain point

Behind any mildly successful woman there is a white man. (No sarcasm. It's true!)

Reason

Stereotype?

Fix (direct)

Break the pattern! Work on a few papers with someone else, or by myself.

Fix (mine)

Jason and I co-founded ML Collective — we are now likely collaborators for life...

- *“You were the middle author in a lot of papers, what exactly were your contributions?”*

Pain point

The increasingly messy credit assignment problem.

Reason

There's now too much (quantifiable) gains associated with ML papers.

Fix (direct)

Break the pattern! Work on a few first-author papers.

Fix (mine)

I want to help others publish first-author or last-author papers.

- *“I can’t quite place you as an expert of anything; what do you want to focus on next?”*

Pain point Yup. I am more of a generalist (in NNs).

Reason The hyperspecialized societal culture has seeped into research.

Fix (direct) Break the pattern! Focus, and make a name in one small thing first.

Fix (mine) I really want to stay curious and open-minded.

ML Collective

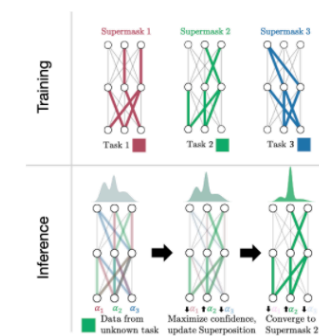
[Home](#) [FAQ](#) [About](#) [Reading Group](#)

ML Collective is an independent, nonprofit organization that conducts fundamental **machine learning research** and provides opportunities for **collaboration** and **mentorship**.

We provide a "research home" to unaffiliated and underrepresented researchers, as well as those on non-traditional paths of AI research. We rely on established researchers who care deeply about the social impact of science, and are eager to give back, broaden and diversify their collaboration circle.

At ML Collective, we believe research opportunities should be **accessible** and **free**, and that **open collaboration** is the key to further democratizing AI research.

Projects

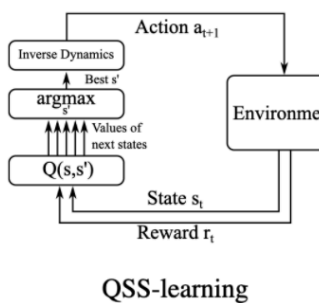


Supermasks in Superposition

Mitchell Wortsman, Vivek Ramanujan, Rosanne Liu, Aniruddha Kembhavi, Mohammad Rastegari, Jason Yosinski, Ali Farhadi

Published at **NeurIPS 2020 (To Appear)**

[arXiv](#) [\(pdf\)](#) | [blog](#) | [code](#)

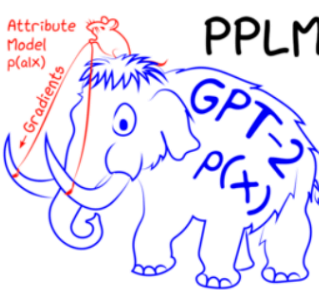


Estimating $Q(s,s')$ with Deep Deterministic Dynamics Gradients

Ashley D. Edwards, Himanshu Sahni, Rosanne Liu, Jane Hung, Ankit Jain, Rui Wang, Adrien Ecoffet, Thomas Miconi, Charles Isbell, Jason Yosinski

Published at **ICML 2020**

[arXiv](#) [\(pdf\)](#) | [code](#) | [video](#)

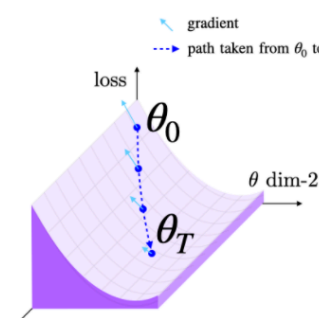


Plug and Play Language Models: a Simple Approach to Controlled Text Generation

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, Rosanne Liu

Published at **ICLR 2020**

[arXiv](#) [\(pdf\)](#) | [blog](#) | [code](#) | [video](#)



LCA: Loss Change Allocation for Neural Network Training

Janice Lan, Rosanne Liu, Hattie Zhou, Jason Yosinski

Published at **NeurIPS 2019**

[arXiv](#) [\(pdf\)](#) | [blog](#) | [code](#) | [video](#)

Members



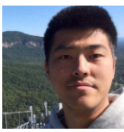
Rosanne Liu

[Home](#) [Blog](#) [Twitter](#) [GitHub](#)



Mitchell Wortsman

[Home](#) [Blog](#) [Twitter](#) [GitHub](#)



Niel Teng Hu

[GitHub](#)



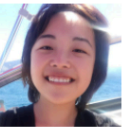
Piero Molino

[Home](#) [Blog](#) [Twitter](#) [GitHub](#)



Xinyu Hu

[Blog](#) [Twitter](#) [GitHub](#)



Hattie Zhou

[Home](#) [Blog](#) [Twitter](#)



Janice Lan

[Blog](#) [Twitter](#)



Rui Wang

[Blog](#) [Twitter](#)



Sam Greydanus

[Home](#) [Blog](#) [Twitter](#) [GitHub](#)



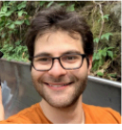
Yariv Sadan

[Twitter](#) [GitHub](#)



Ashley Edwards

[Home](#) [Blog](#) [Twitter](#) [GitHub](#)



Eric Frank

[Blog](#) [Twitter](#)



Thomas Miconi

[Blog](#) [Twitter](#)



Andrea Madotto

[Home](#) [Blog](#) [Twitter](#) [GitHub](#)



Jane Hung

[Blog](#) [Twitter](#) [GitHub](#)



Ankit Jain

[Blog](#) [Twitter](#)



Sara Hooker

[Home](#) [Blog](#) [Twitter](#) [GitHub](#)



Stephanie Sher

[Twitter](#)



Sebastian Ruder

[Home](#) [Blog](#) [Twitter](#) [GitHub](#)



Jonathan Frankle

[Home](#) [Blog](#) [Twitter](#)



Chloe Hsu

[Home](#) [Blog](#) [Twitter](#) [GitHub](#)



Chirag Agarwal

[Home](#) [Blog](#) [Twitter](#) [GitHub](#)



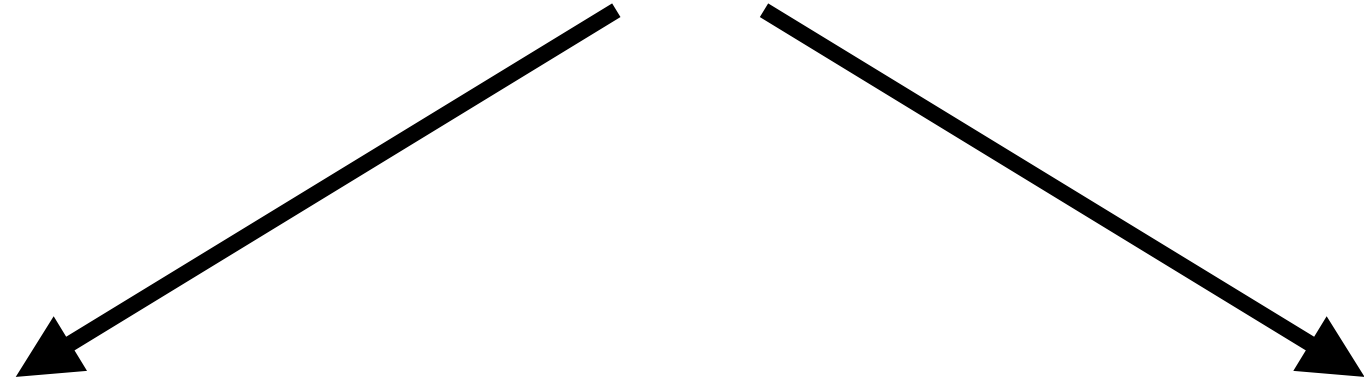
Jason Yosinski

[Home](#) [Blog](#) [Twitter](#) [GitHub](#)

<https://mlcollective.org>

JOB ?

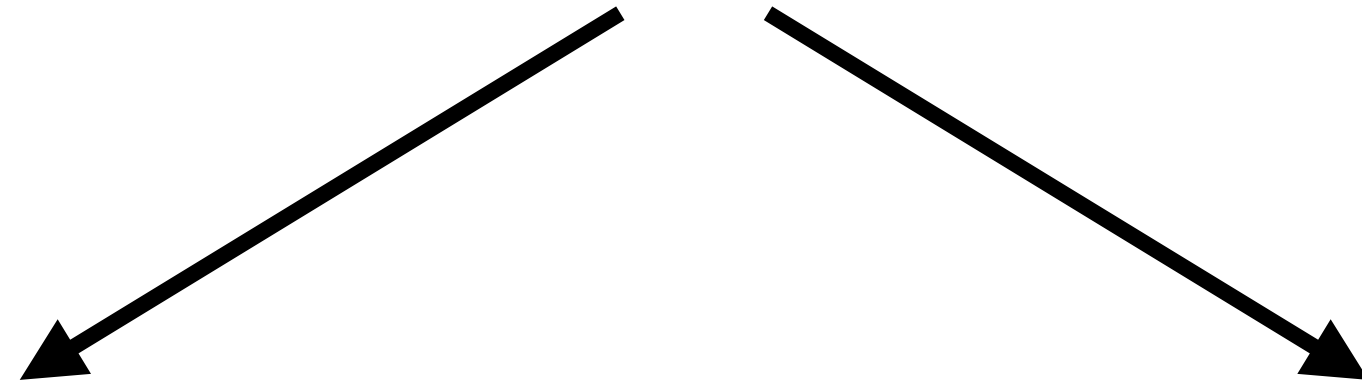
JOB



Pays

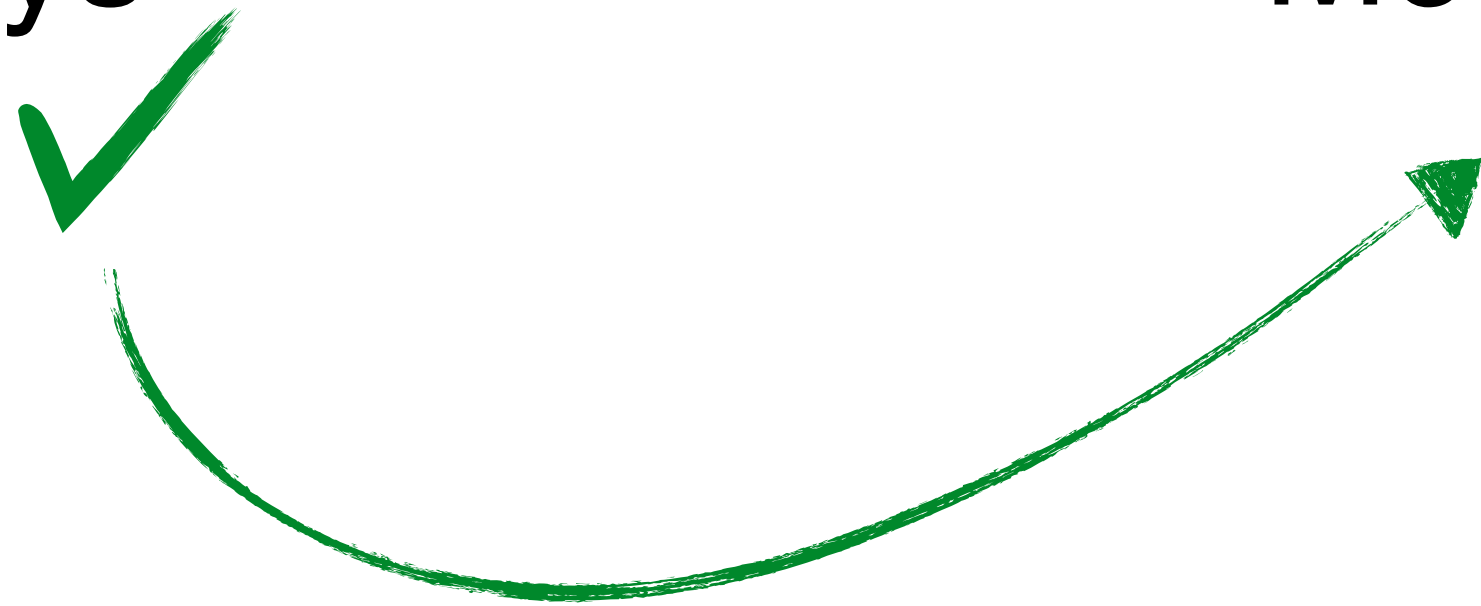
Meaningful

JOB

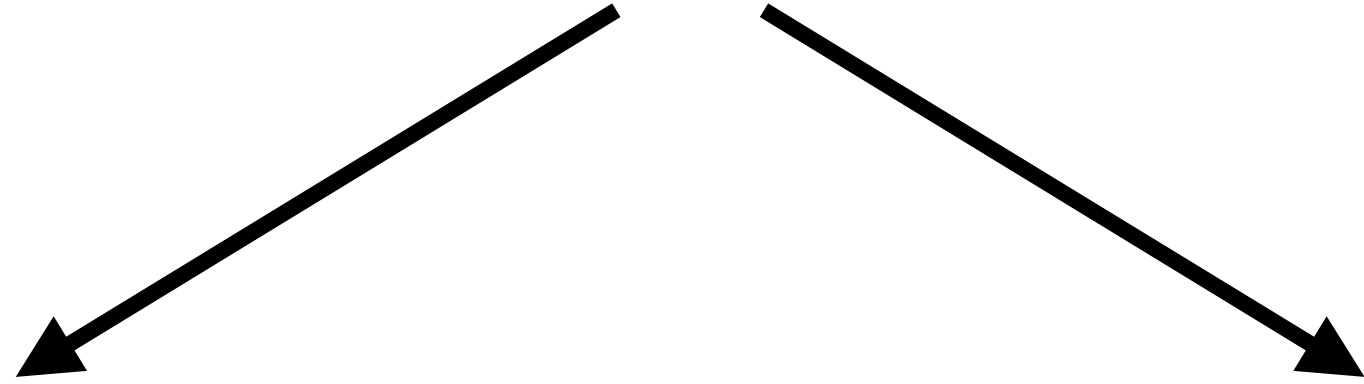


Pays

Meaningful

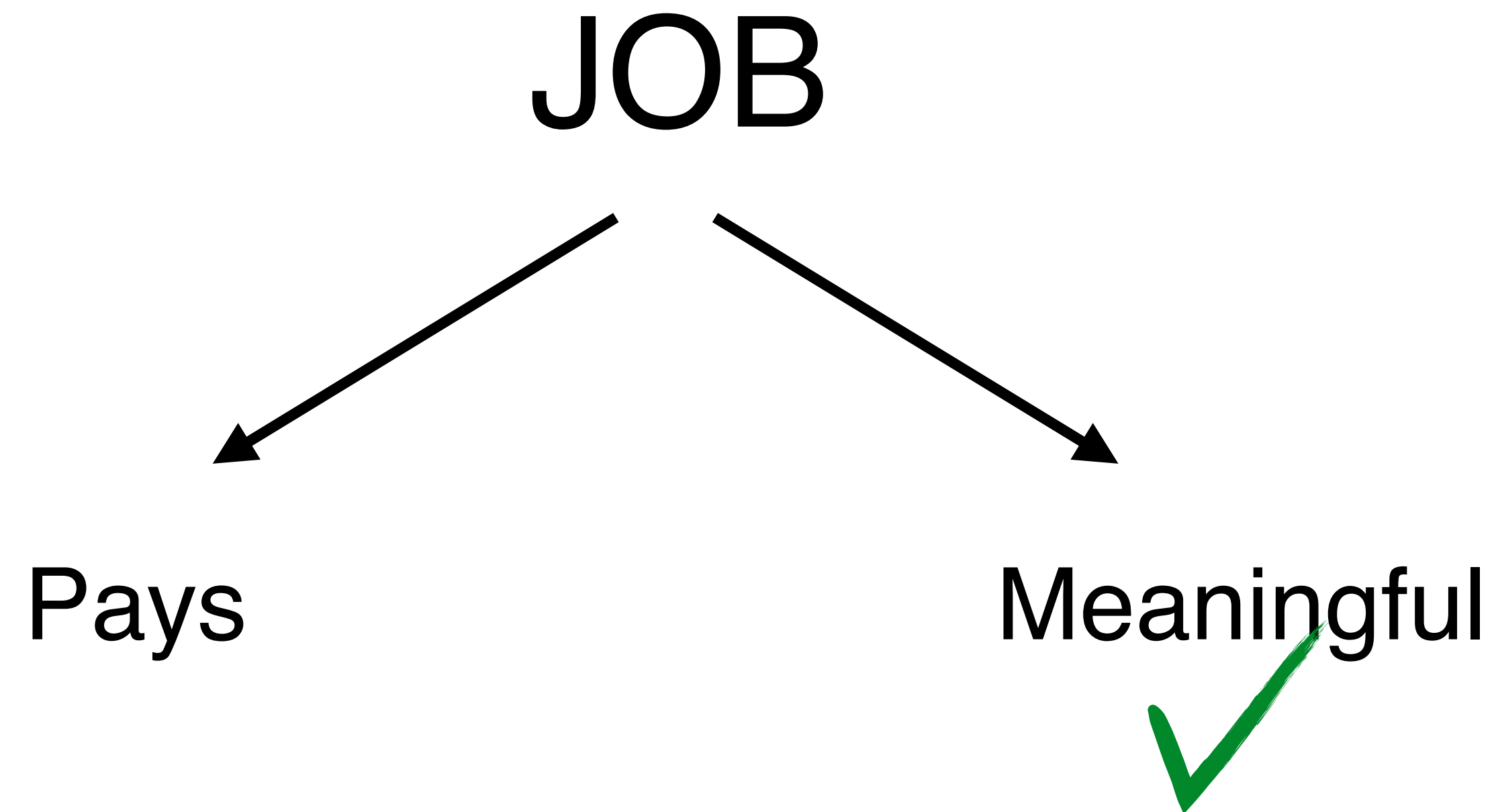


JOB



Pays

Meaningful



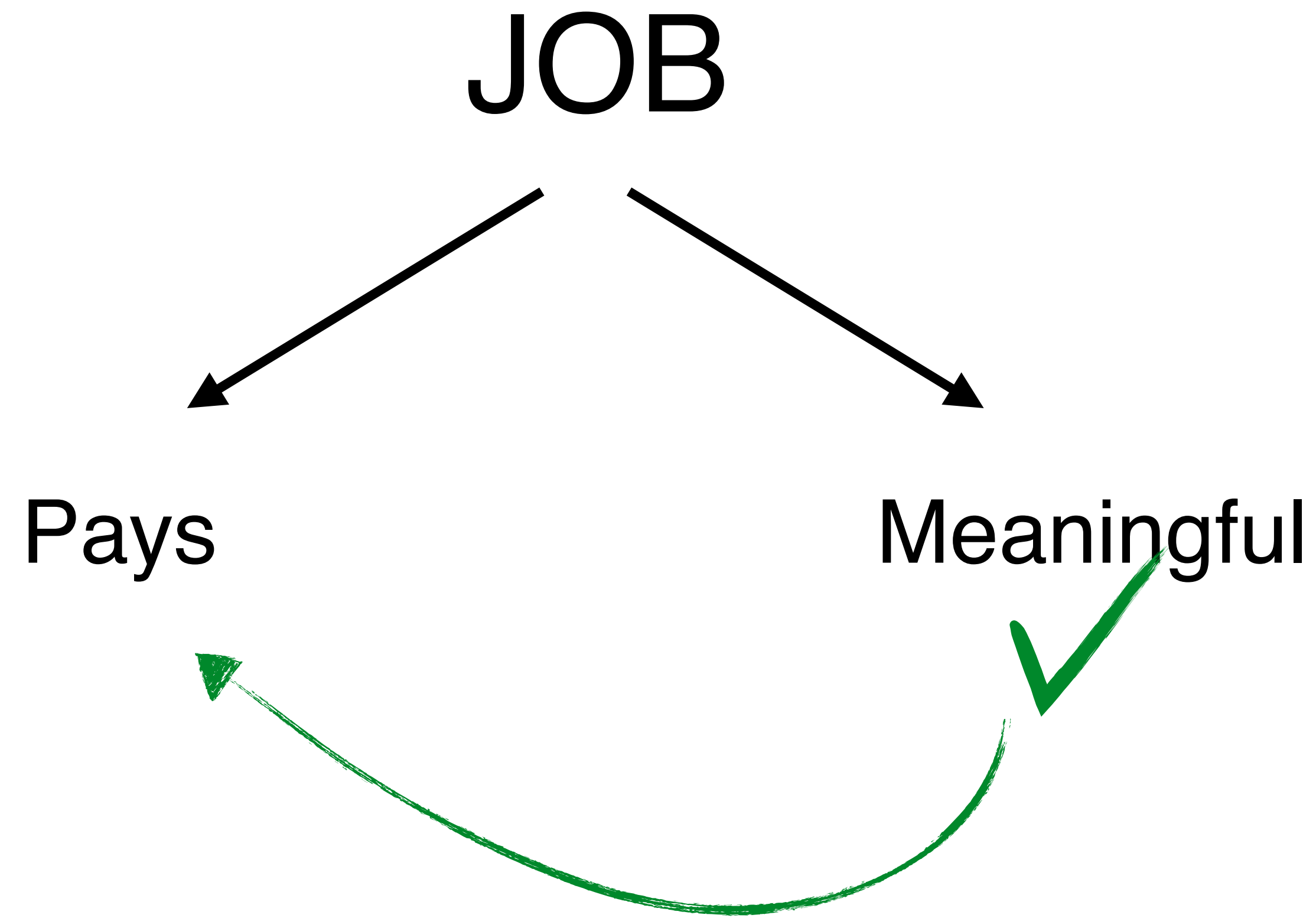
Help people in a similar misery — help them publish their 1st 1st-author paper (or last-author paper), help connect them to the right mentor, get into ML, know what to expect & what to do...

Help normalize the expectation of ML research

Help level the playing field by redistributing opportunity

Help build a community of collaborators, not competitors

Help widen the path, at least a little bit



Help people in a similar misery — help them publish their 1st 1st-author paper (or last-author paper), help connect them to the right mentor, get into ML, know what to expect & what to do...

Help normalize the expectation of ML research

Help level the playing field by redistributing opportunity

Help build a community of collaborators, not competitors

Help widen the path, at least a little bit

Misconception #4: Goal-driven vs. Curiosity-driven research

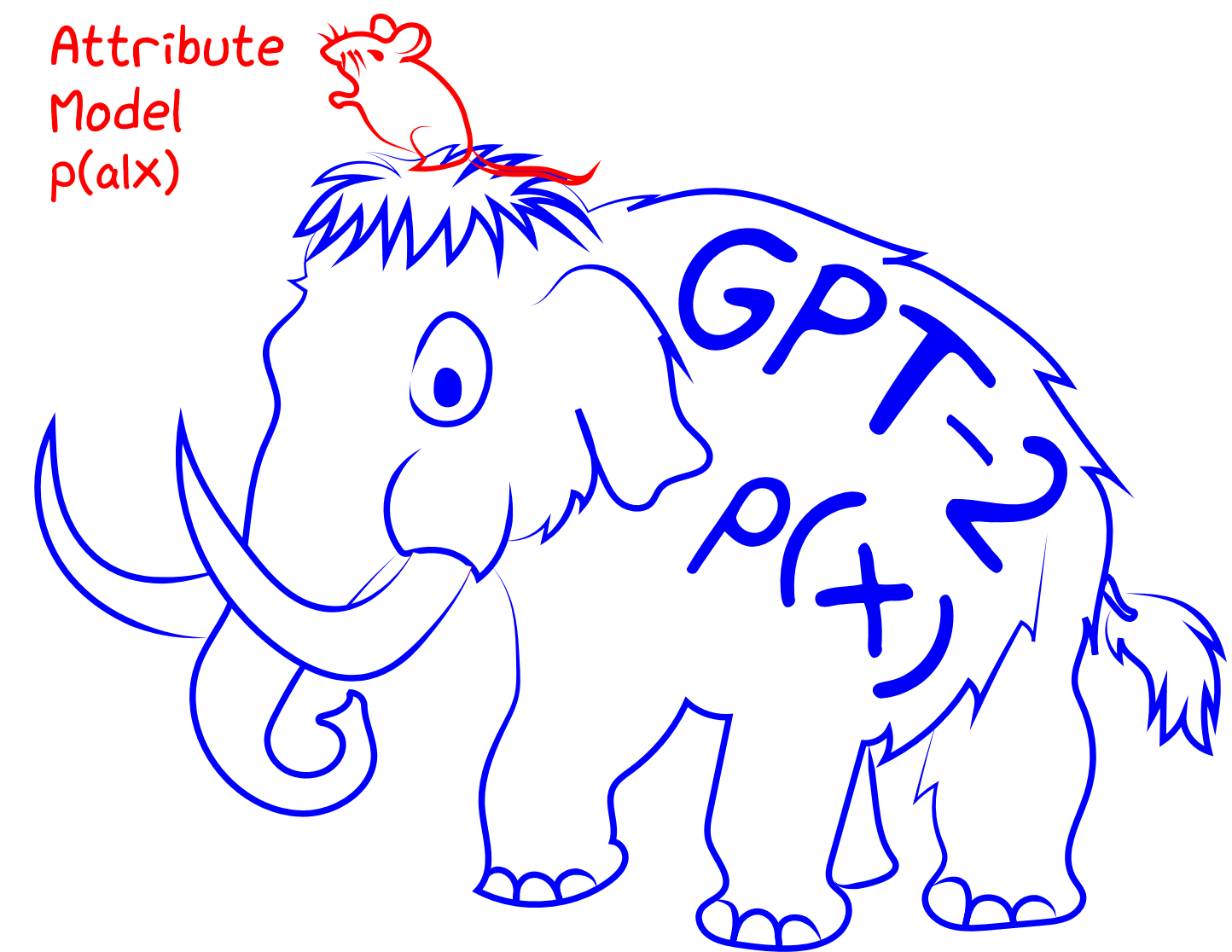
Misconception #5: fast vs. slow science

Let's spend some time diving in this largely curiosity-driven research we did with large language models — GPT-2.

Plug and Play Language Models



Plug and Play Language Models



Plug and Play Language Models

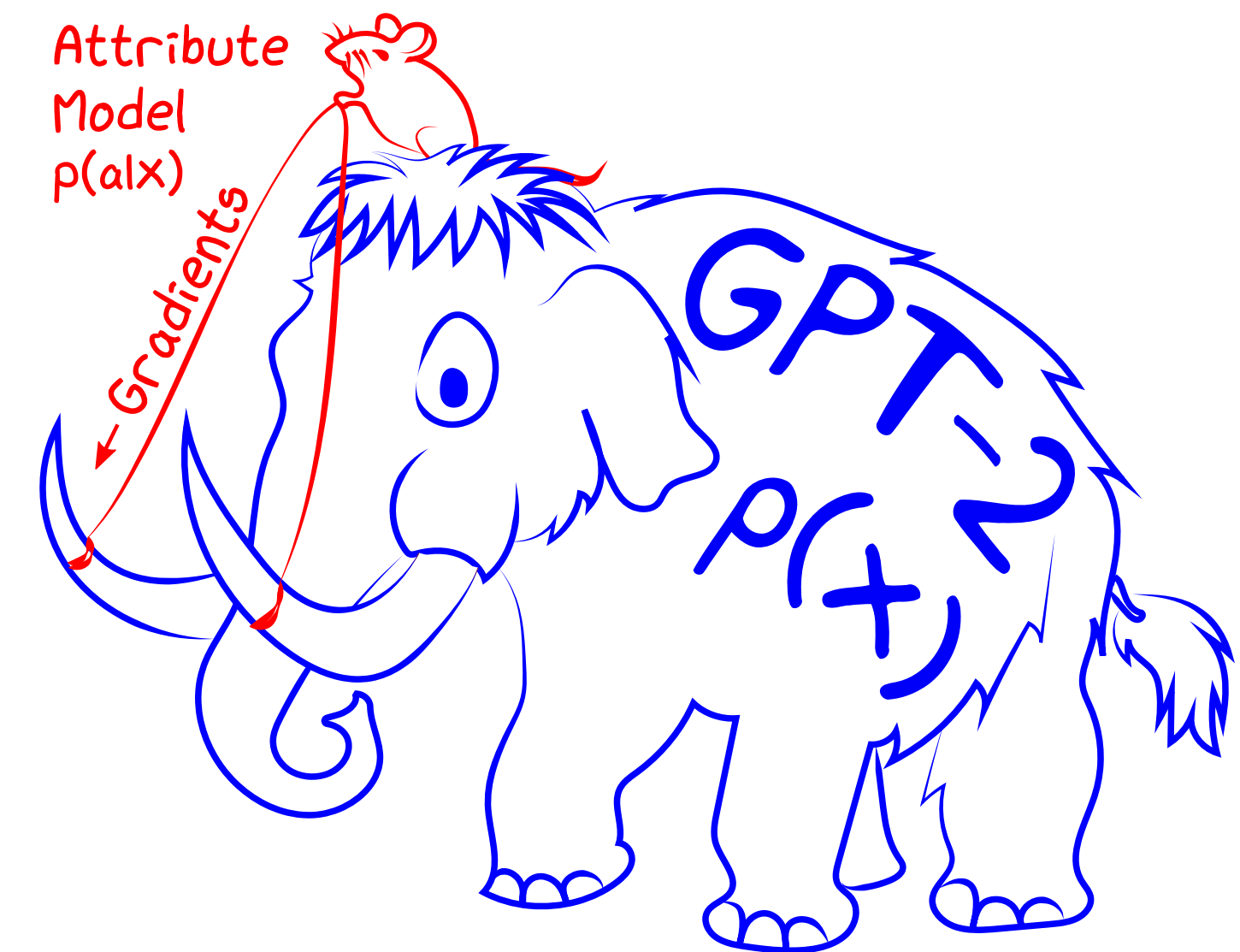
[-] The potato is a plant from the family of the same name that can be used as a condiment and eaten raw. It can also be eaten raw in its natural state...



Plug and Play Language Models

[-] *The potato* is a plant from the family of the same name that can be used as a condiment and eaten raw. It can also be eaten raw in its natural state...

[Negative] *The potato* is a pretty bad idea. It can make you fat, it can cause you to have a terrible immune system, and it can even kill you...

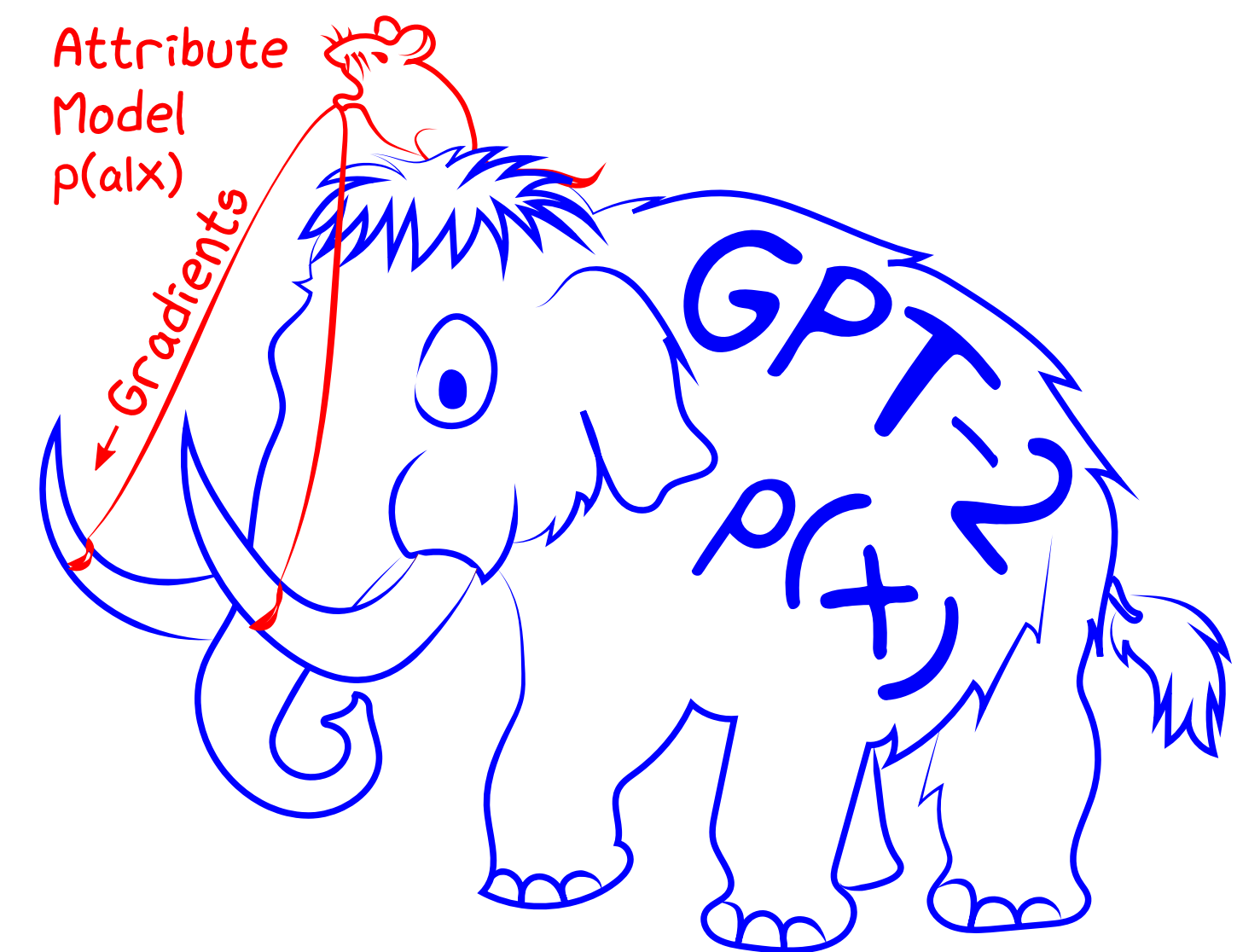


Plug and Play Language Models

[-] *The potato* is a plant from the family of the same name that can be used as a condiment and eaten raw. It can also be eaten raw in its natural state...

[Negative] *The potato* is a pretty bad idea. It can make you fat, it can cause you to have a terrible immune system, and it can even kill you...

[Positive] *The potato* chip recipe you asked for! We love making these, and I've been doing so for years. I've always had a hard time keeping a recipe secret. I think it's the way our kids love to eat them...



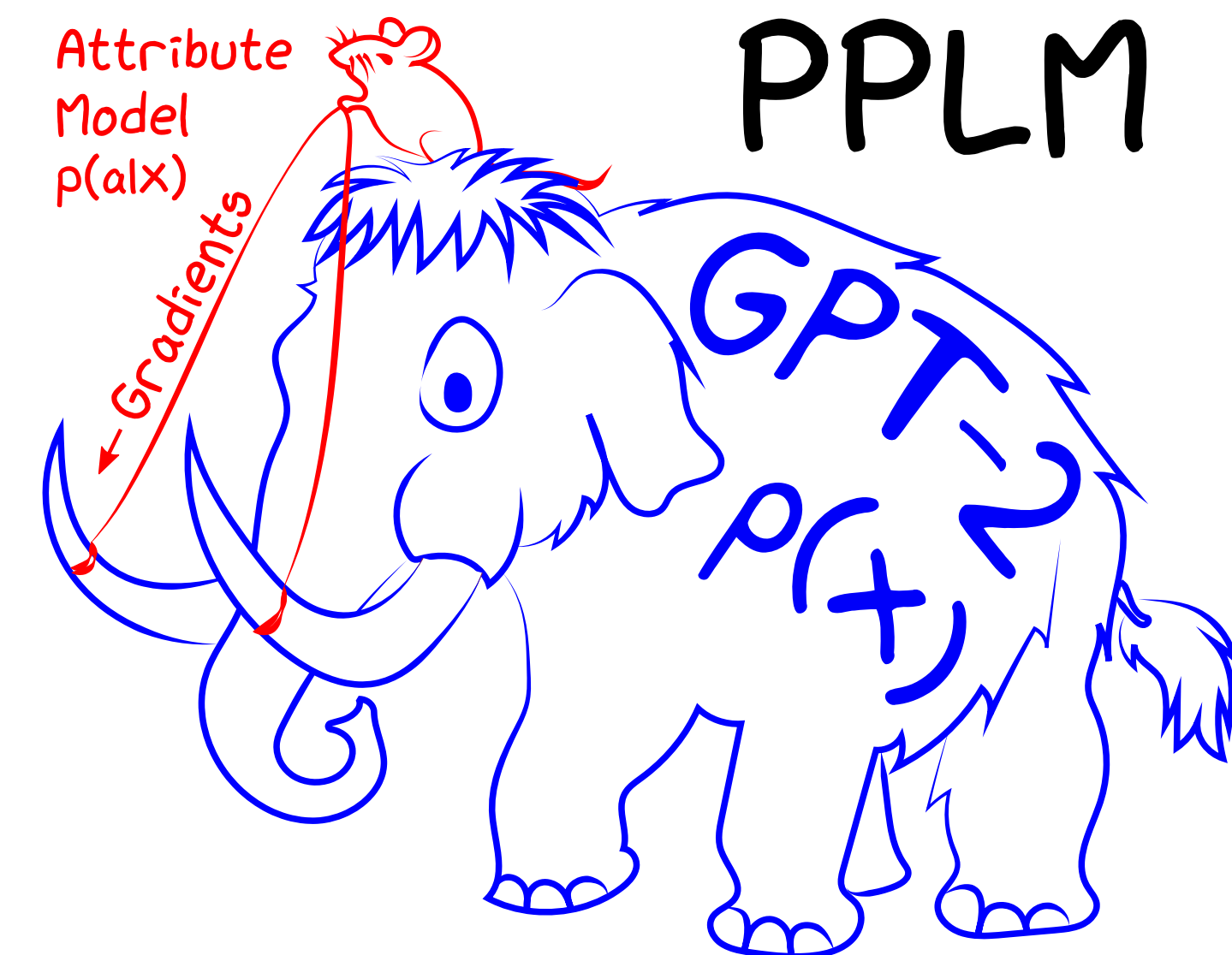


Plug and Play Language Models: A Simple Approach Towards Controlled Text Generation

[-] The potato is a plant from the family of the same name that can be used as a condiment and eaten raw. It can also be eaten raw in its natural state...

[Negative] The potato is a pretty bad idea. It can make you fat, it can cause you to have a terrible immune system, and it can even kill you...

[Positive] The potato chip recipe you asked for! We love making these, and I've been doing so for years. I've always had a hard time keeping a recipe secret. I think it's the way our kids love to eat them...



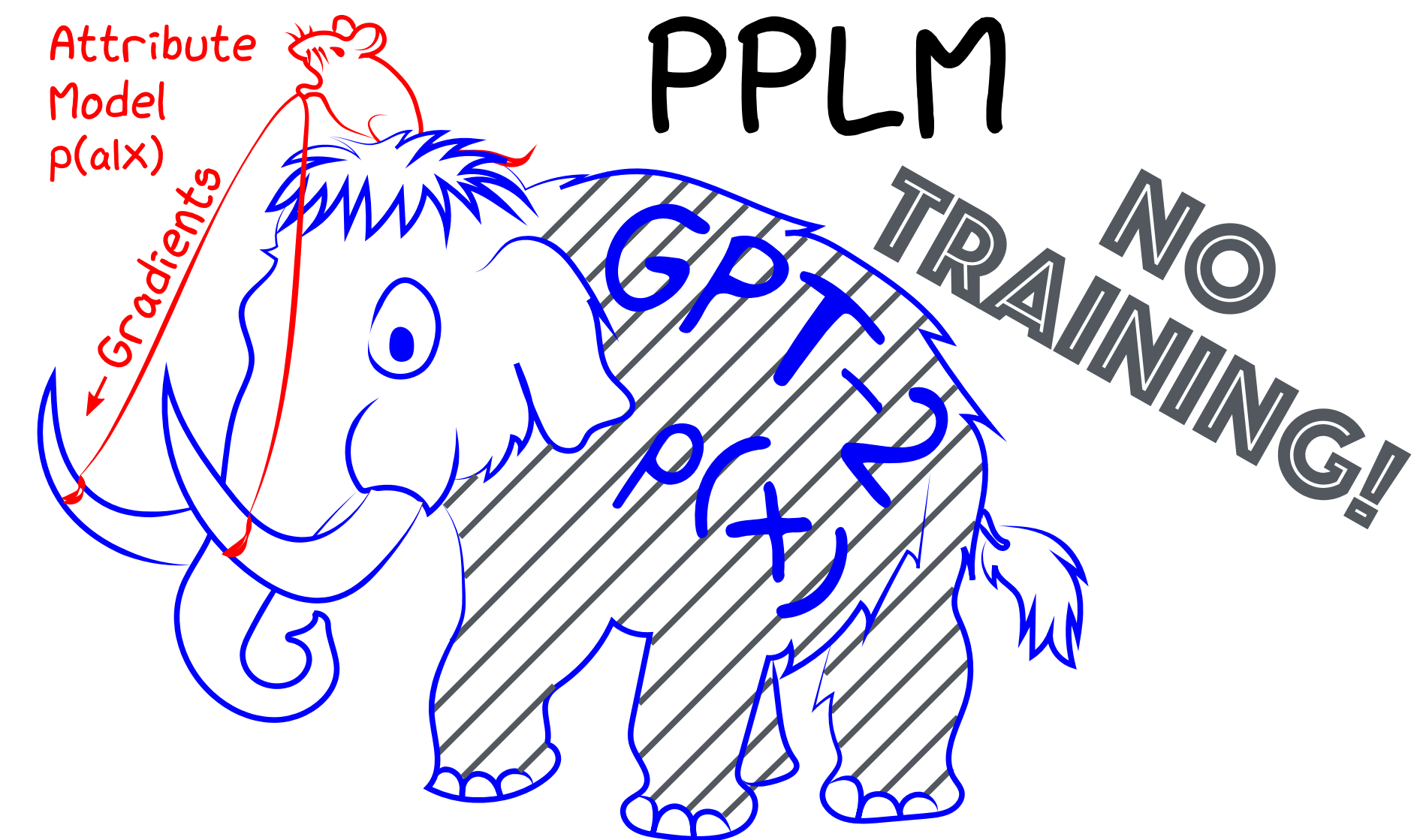


Plug and Play Language Models: A Simple Approach Towards Controlled Text Generation

[-] The potato is a plant from the family of the same name that can be used as a condiment and eaten raw. It can also be eaten raw in its natural state...

[Negative] The potato is a pretty bad idea. It can make you fat, it can cause you to have a terrible immune system, and it can even kill you...

[Positive] The potato chip recipe you asked for! We love making these, and I've been doing so for years. I've always had a hard time keeping a recipe secret. I think it's the way our kids love to eat them...



Transformer Language Models (Google, OpenAI)

arXiv:1706.03762v5 [cs.CL] 6 Dec 2017

Attention Is All You Need

Ashish Vaswani*
Google Brain
avaswani@google.com

Noam Shazeer*
Google Brain
noam@google.com

Niki Parmar*
Google Research
nikip@google.com

Jakob Uszkoreit*
Google Research
usz@google.com

Llion Jones*
Google Research
llion@google.com

Aidan N. Gomez*[†]
University of Toronto
aidan@cs.toronto.edu

Lukasz Kaiser*
Google Brain
lukaszkaiser@google.com

Illia Polosukhin*[‡]
illia.polosukhin@gmail.com

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

1 Introduction


Recurrent neural networks, long short-term memory [13] and gated recurrent [7] neural networks in particular, have been firmly established as state of the art approaches in sequence modeling and

*Equal contribution. Listing order is random. Jakob proposed replacing RNNs with self-attention and started the effort to evaluate this idea. Ashish, with Illia, designed and implemented the first Transformer models and has been crucially involved in every aspect of this work. Noam proposed scaled dot-product attention, multi-head attention and the parameter-free position representation and became the other person involved in nearly every detail. Niki designed, implemented, tuned and evaluated countless model variants in our original codebase and tensor2tensor. Llion also experimented with novel model variants, was responsible for our initial codebase, and efficient inference and visualizations. Lukasz and Aidan spent countless long days designing various parts of and implementing tensor2tensor, replacing our earlier codebase, greatly improving results and massively accelerating our research.


[†]Work performed while at Google Brain.
[‡]Work performed while at Google Research.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

(Vaswani et al. 2017)



ABOUTPROGRESSRESOURCESBLOG



FEBRUARY 14, 2019

24 MINUTE READ

Better Language Models and Their Implications

We've trained a large-scale unsupervised language model which generates coherent paragraphs of text, achieves state-of-the-art performance on many language modeling benchmarks, and performs rudimentary reading comprehension, machine translation, question answering, and summarization—all without task-specific training.

VIEW CODE

READ PAPER

Our model, called GPT-2 (a successor to [GPT](#)), was trained simply to predict the next word in 40GB of Internet text. Due to our concerns about malicious applications of the technology, we are not releasing the trained model. As an experiment in responsible disclosure, we are instead releasing a much [smaller model](#) for researchers to experiment with, as well as a [technical paper](#).

(Radford et al. 2019)

Language Modeling (OpenAI)

Human Prompt

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.

Machine
Completion

(Radford et al. 2019)

Language Modeling (OpenAI)

Human Prompt

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.

Machine
Completion

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.

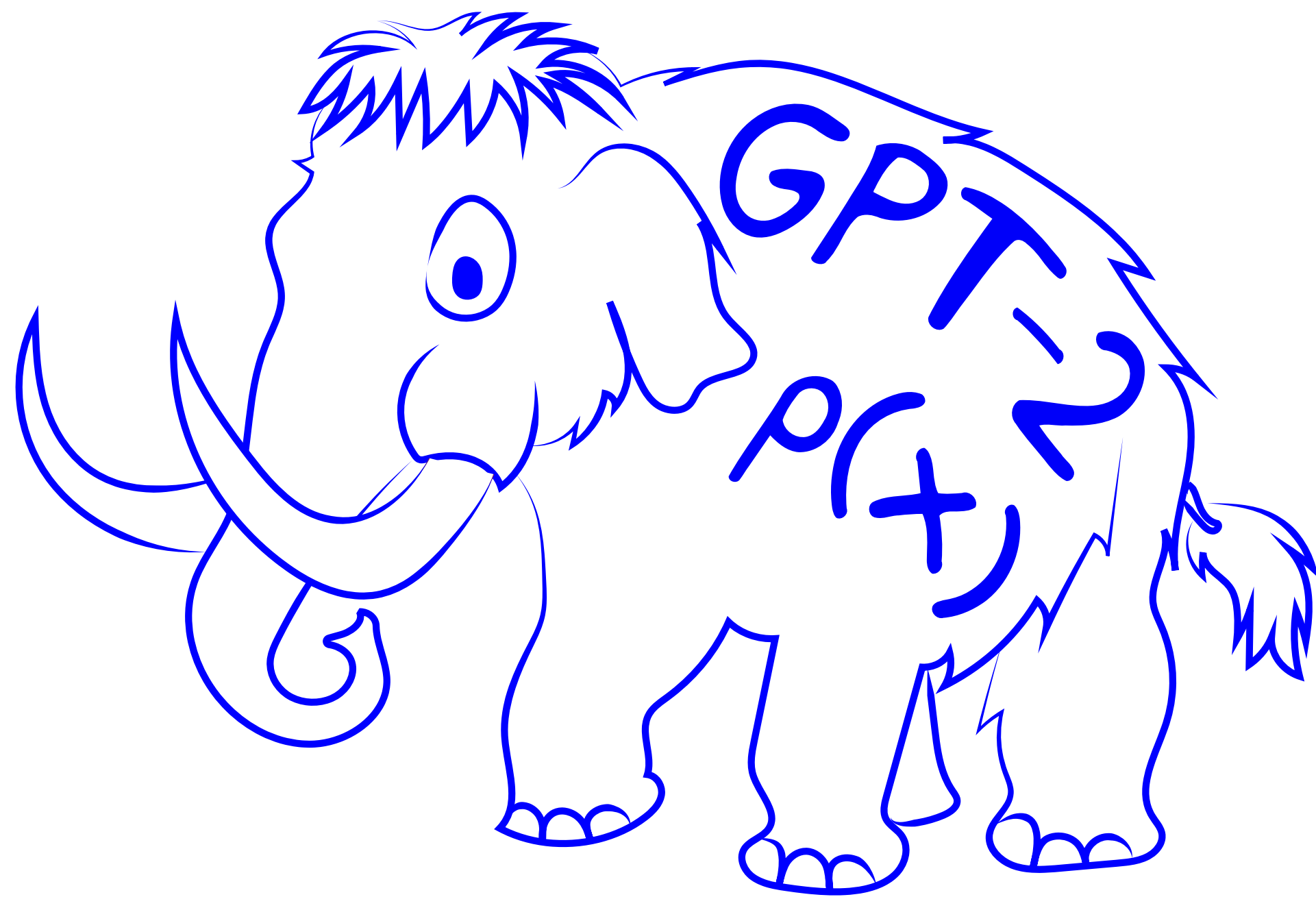
...

(Radford et al. 2019)

Large (Uncontrolled) Language Models $p(x)$

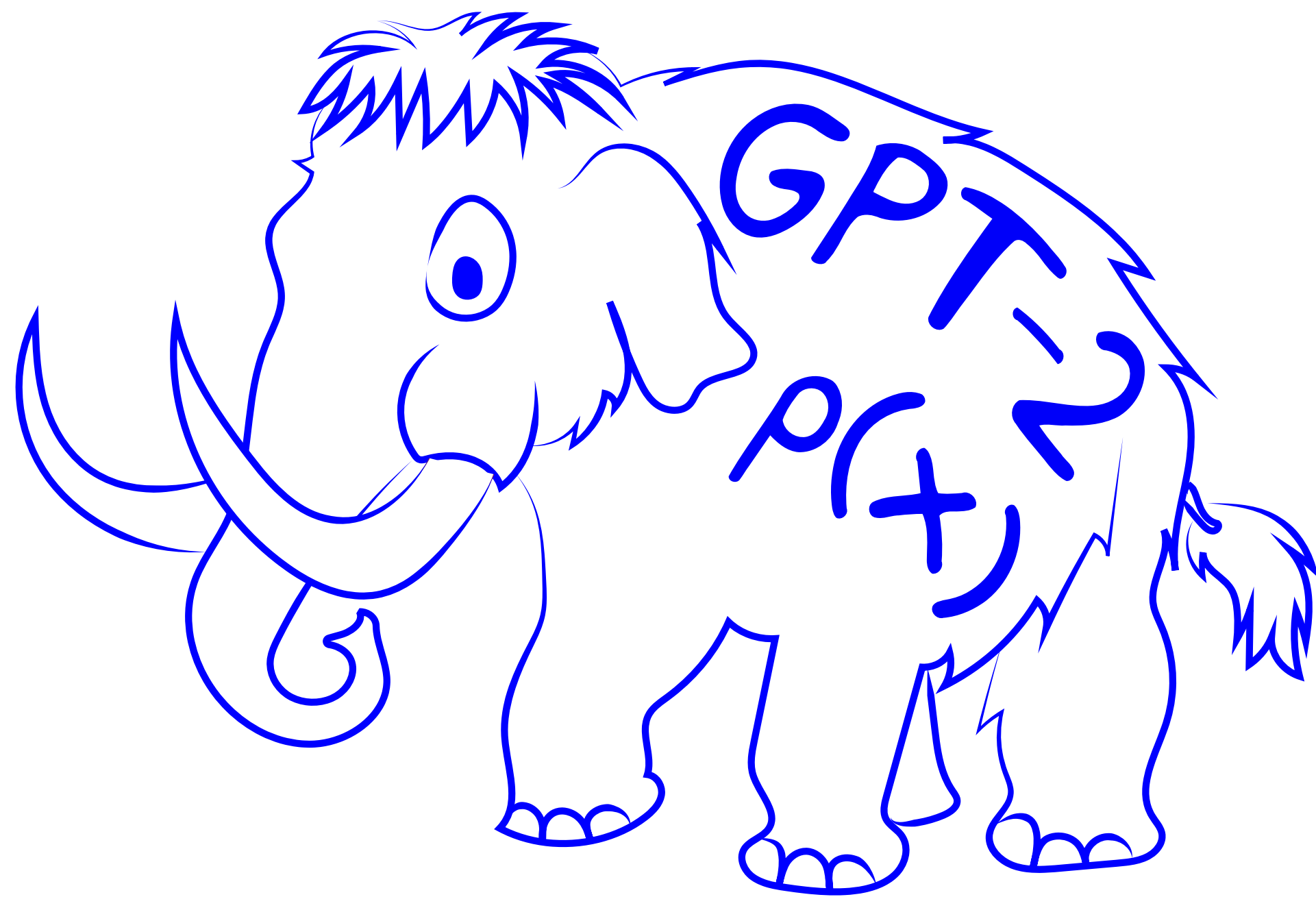


Large (Uncontrolled) Language Models $p(x)$



[-] *The potato* is a plant from the family of the same name that can be used as a condiment and eaten raw. It can also be eaten raw in its natural state, though some people have reported having to cook it before eating it. Its seeds are bitter ...

Controlled Generation $p(x|a)$

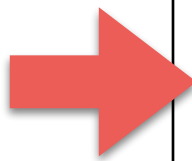


+

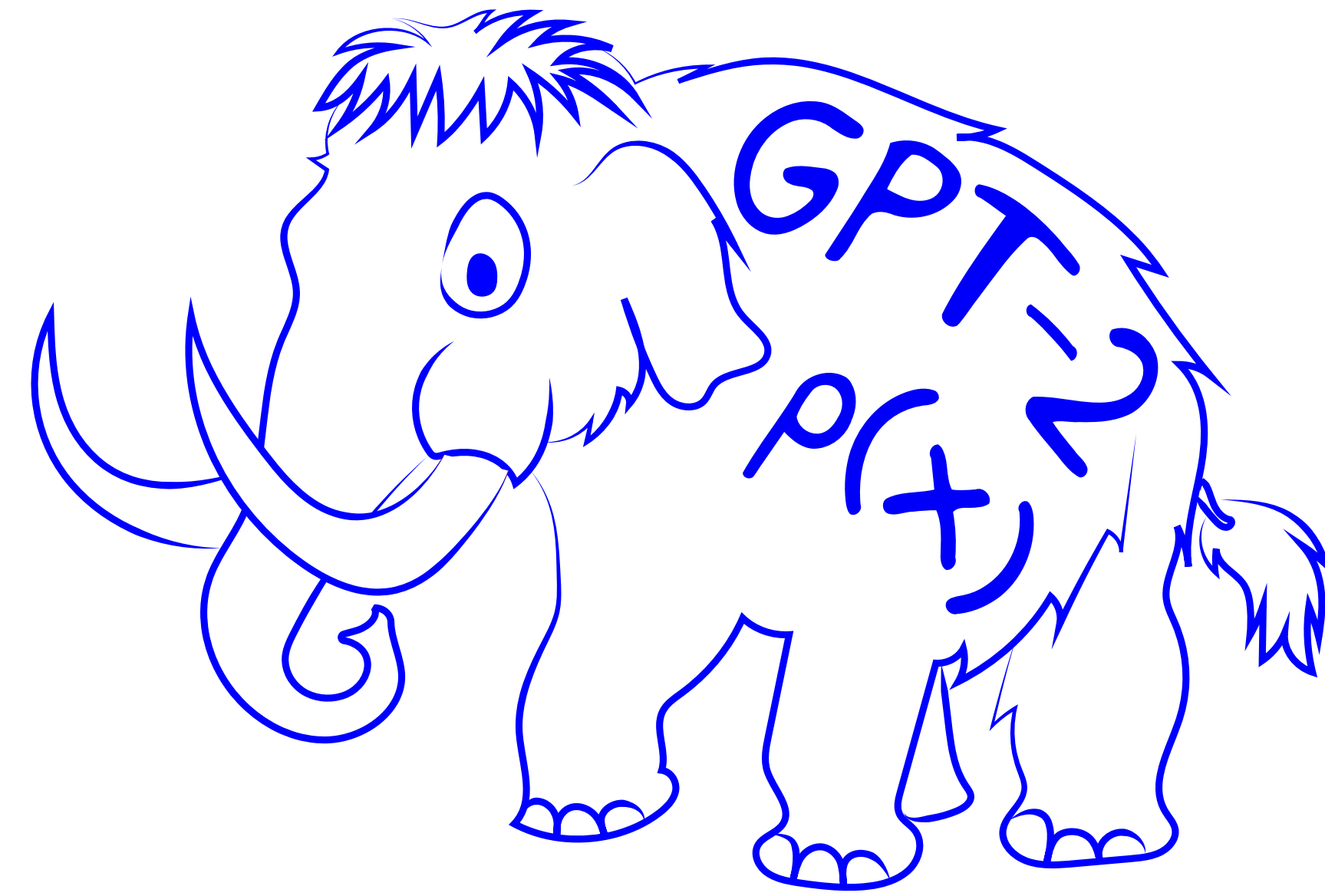


Model Type	Form of Model	Example models and number of trainable parameters
Fine-Tuned Language Model	$p (x)$	Fine-Tuned GPT-2 medium: 345M (Ziegler et al., 2019)
Conditional Language Model	$p (x \mid a)$	CTRL: 1.6B (Keskar et al., 2019)

Model Type	Form of Model	Example models and number of trainable parameters
Fine-Tuned Language Model	$p(x)$	Fine-Tuned GPT-2 medium: 345M (Ziegler et al., 2019)
Conditional Language Model	$p(x a)$	CTRL: 1.6B (Keskar et al., 2019)
Plug and Play Language Model (PPLM)	$p(x a) \propto p(x) p(a x)$ Difficult Easy	PPLM-BoW: 0 (curated word list) PPLM-Discrim: 1K/attribute (not counting pretraining $p(x)$)

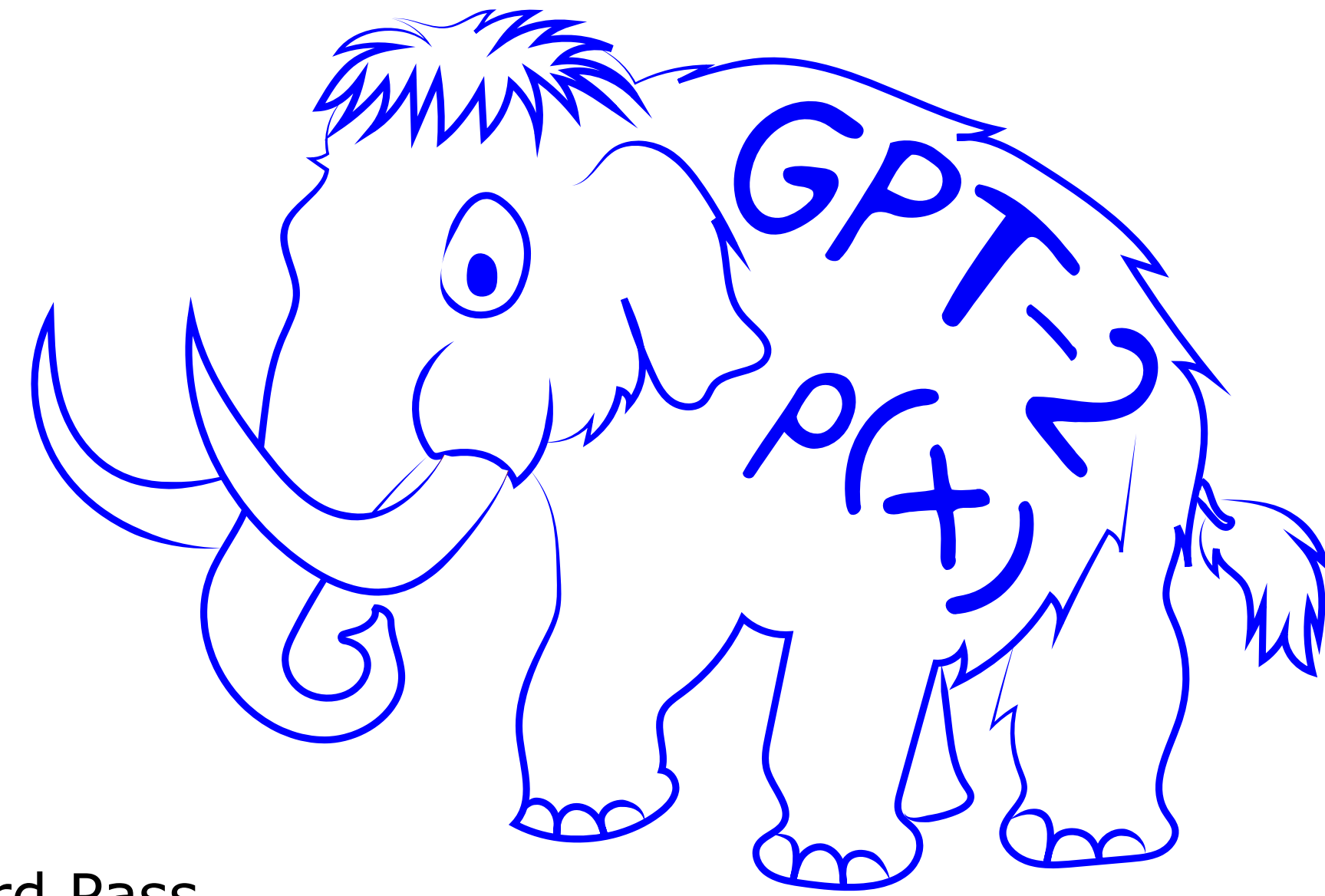


Approach: Ascending $\log p(a|x)$



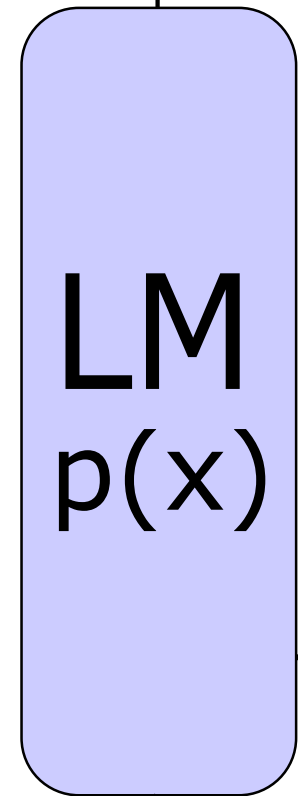
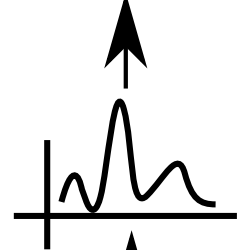
LM
 $p(x)$

Approach: Ascending $\log p(a|x)$



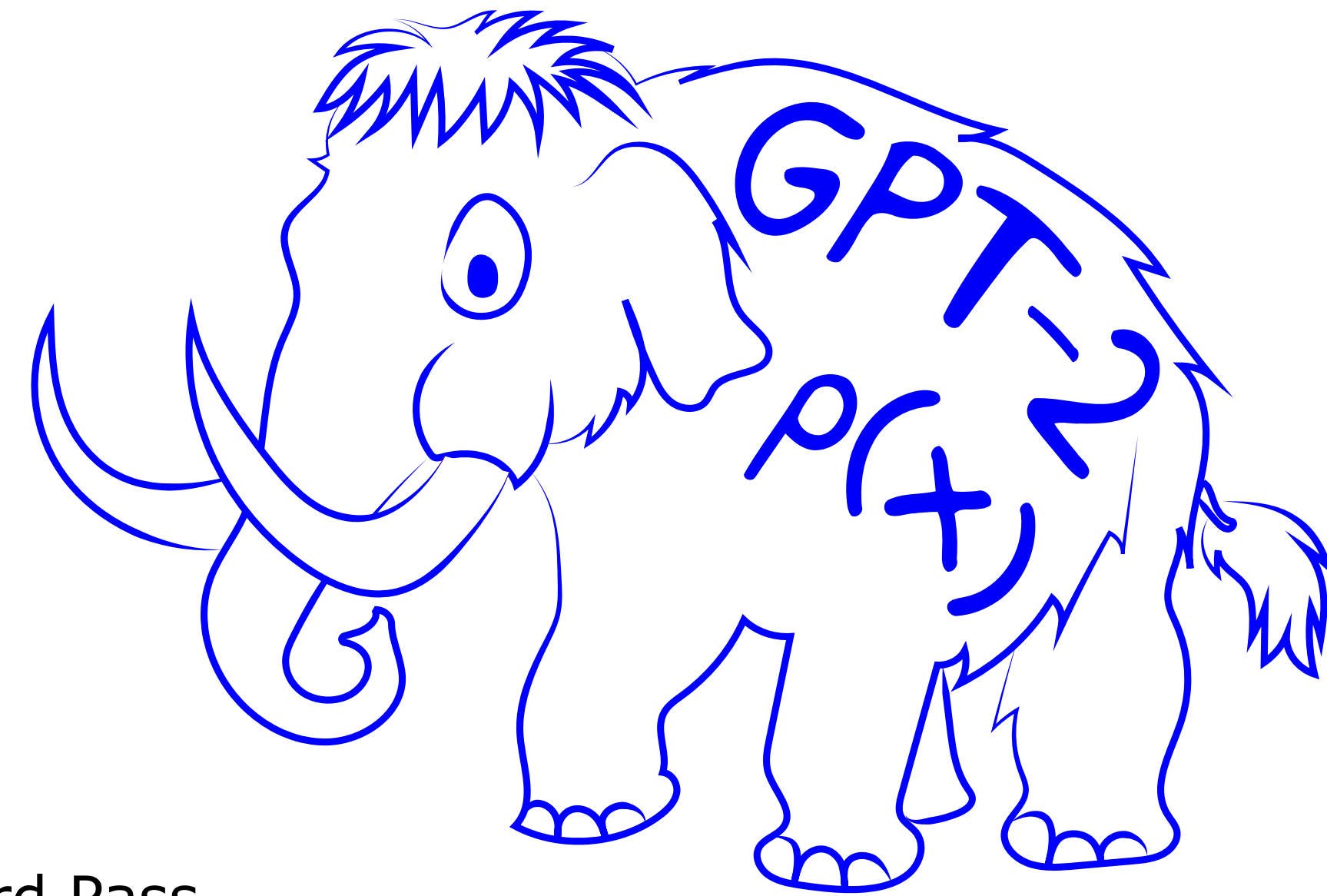
→ Forward Pass

chicken

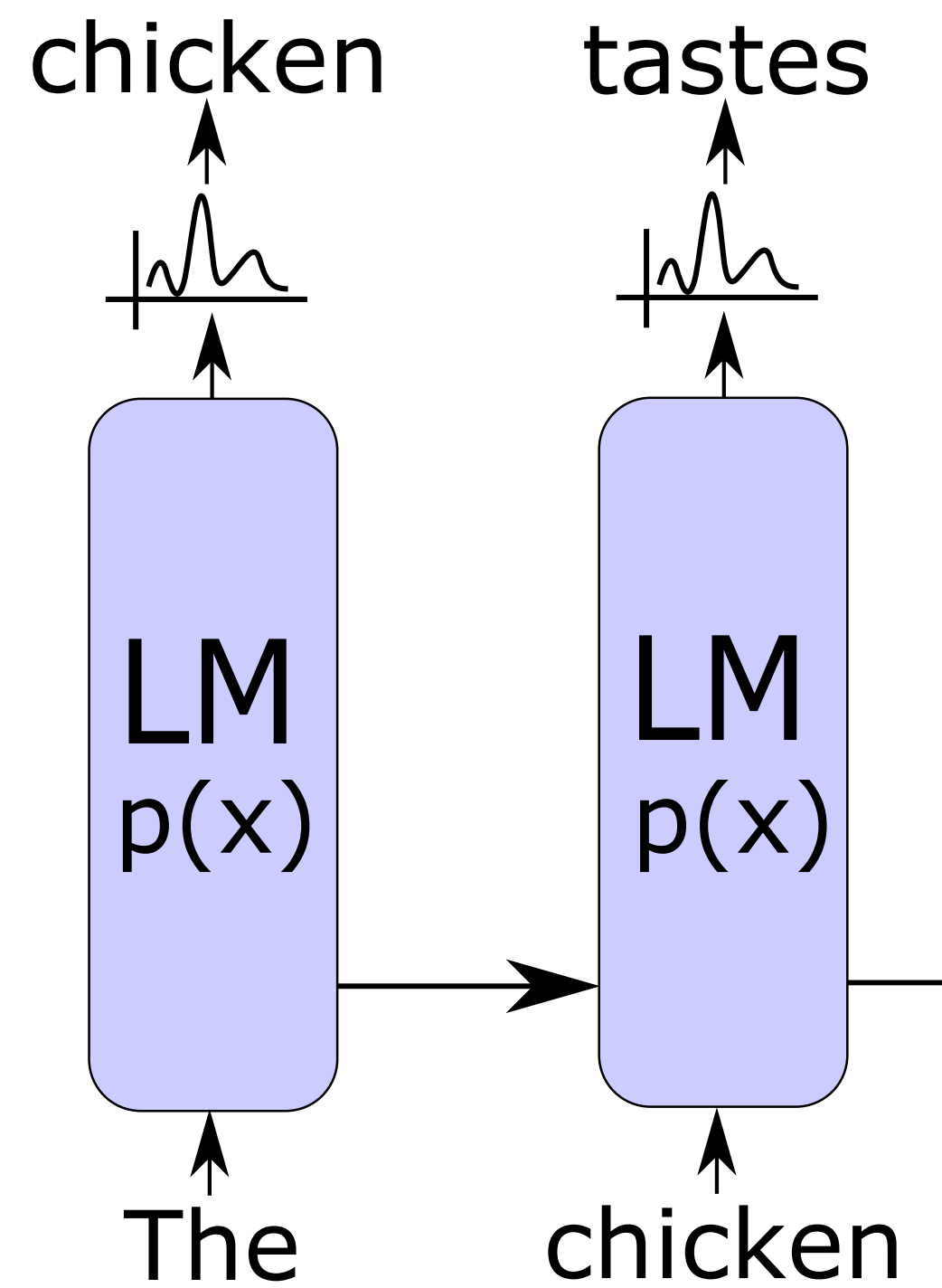


The

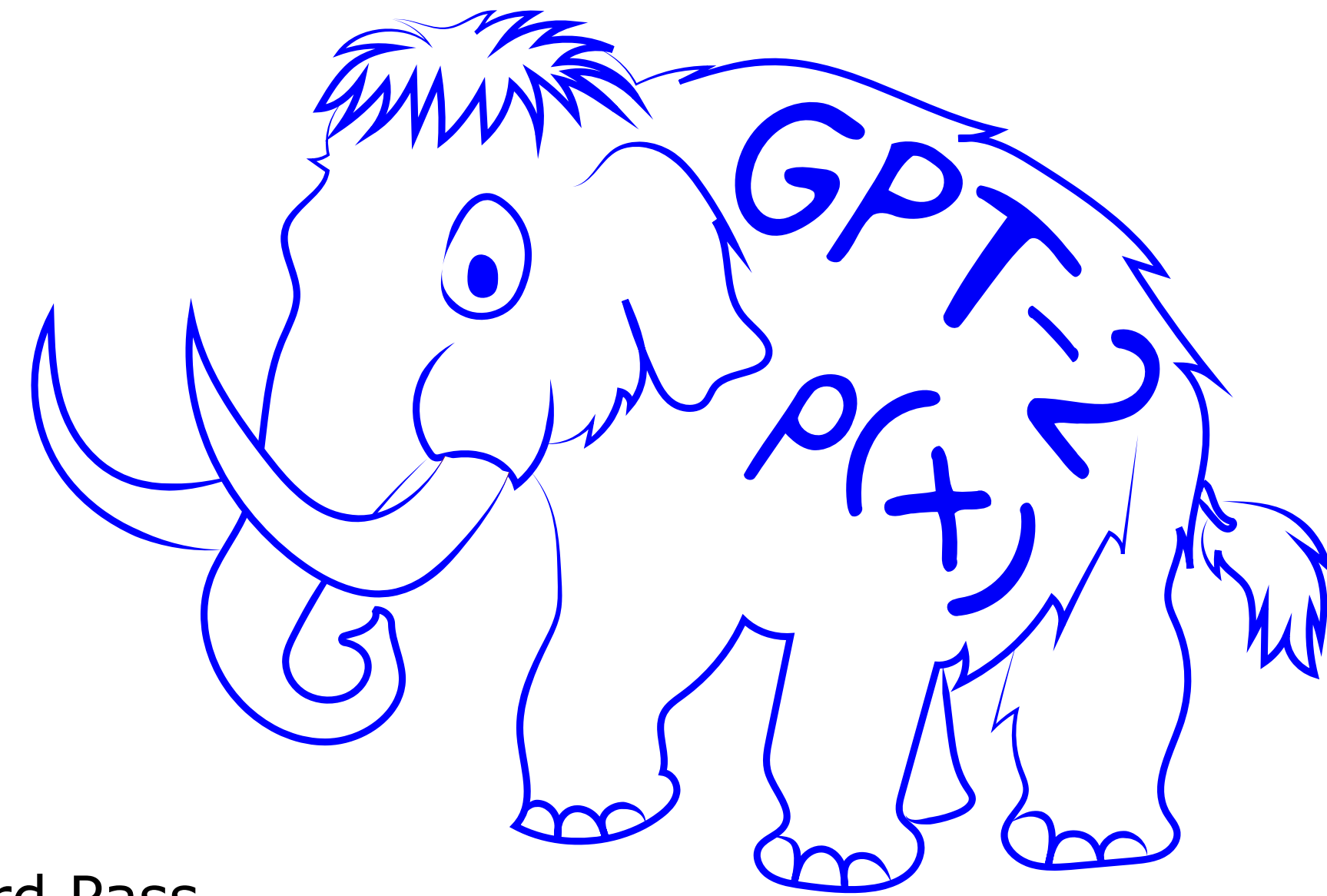
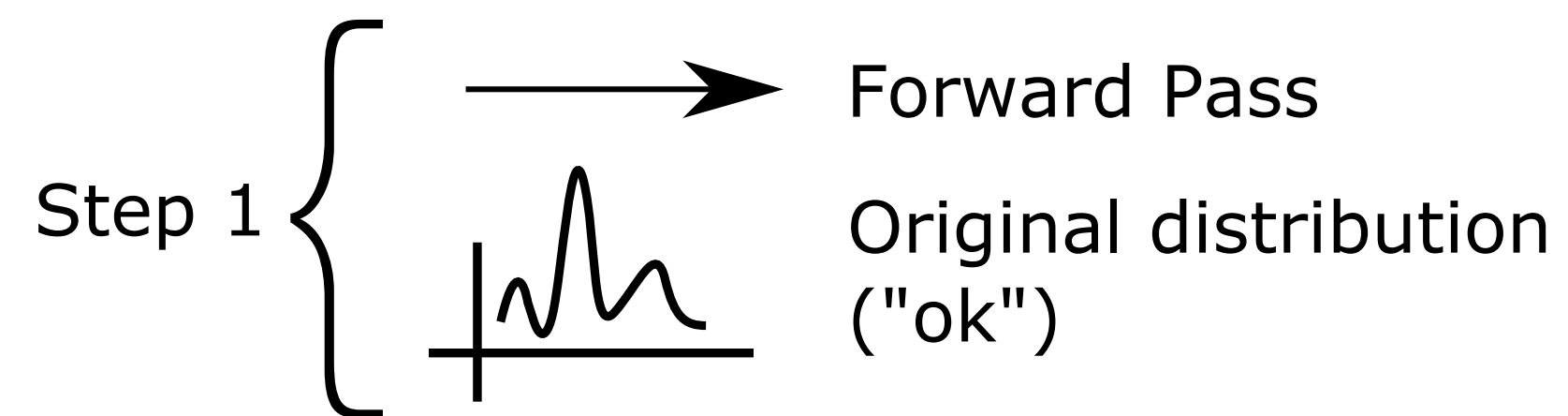
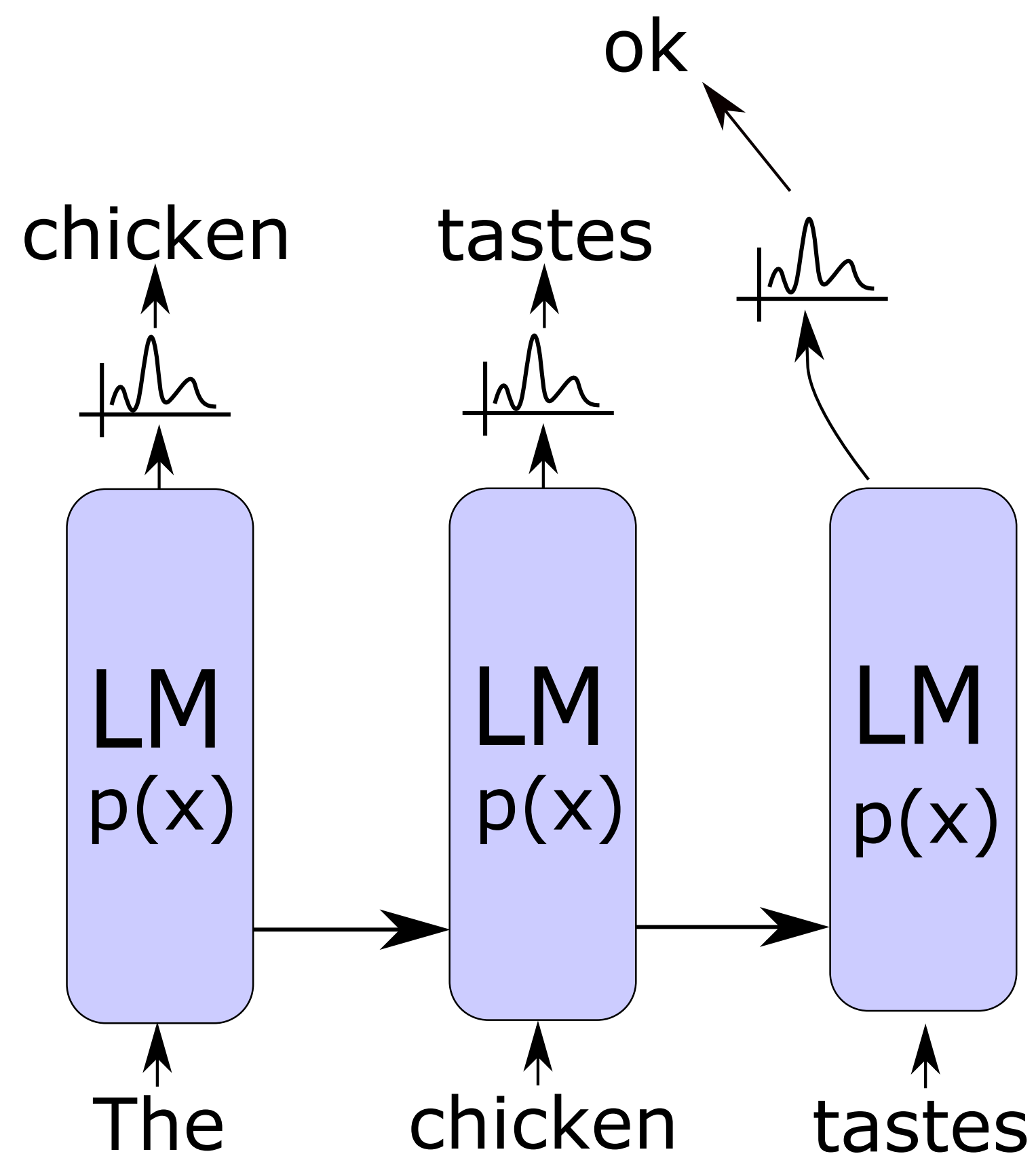
Approach: Ascending $\log p(a|x)$



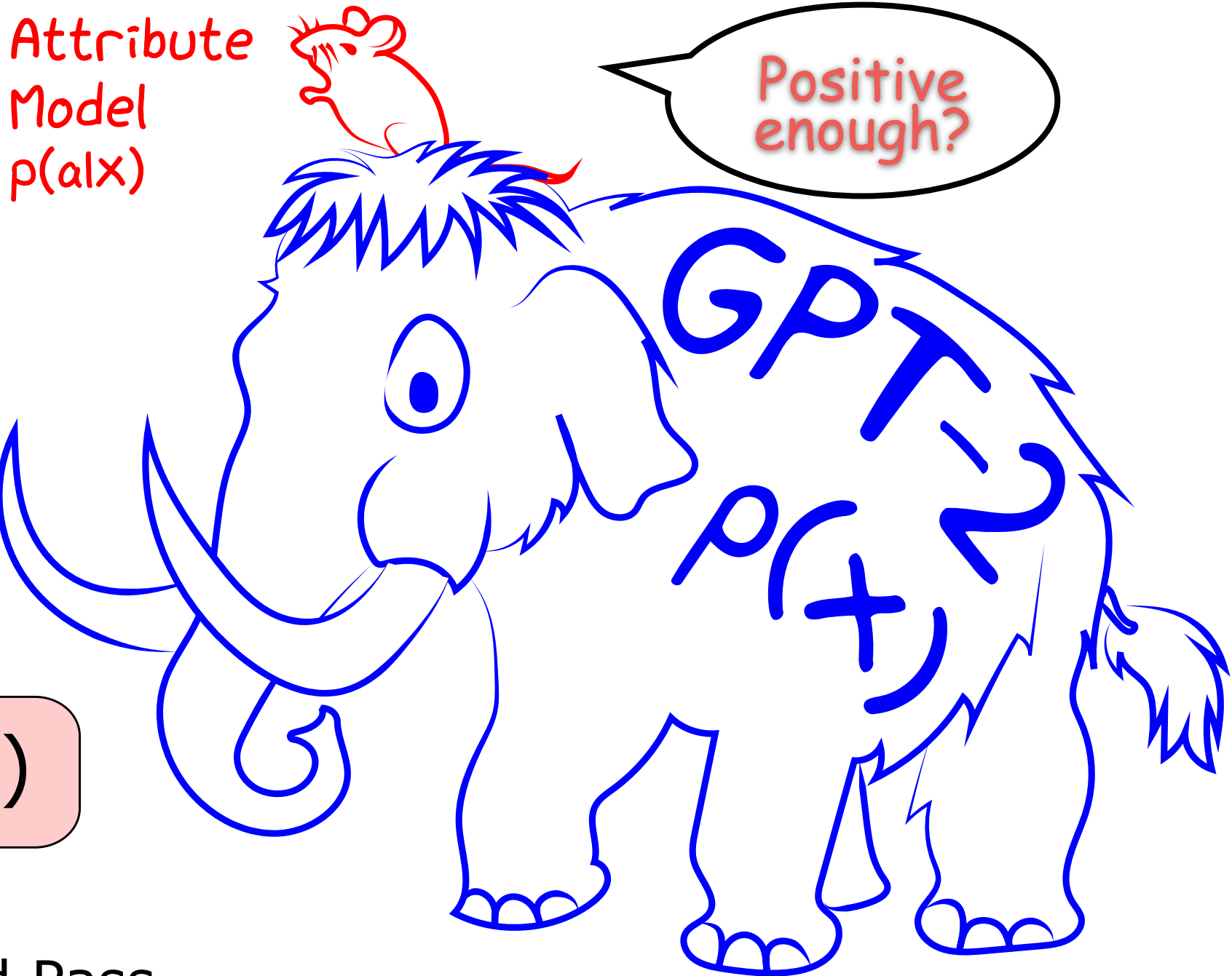
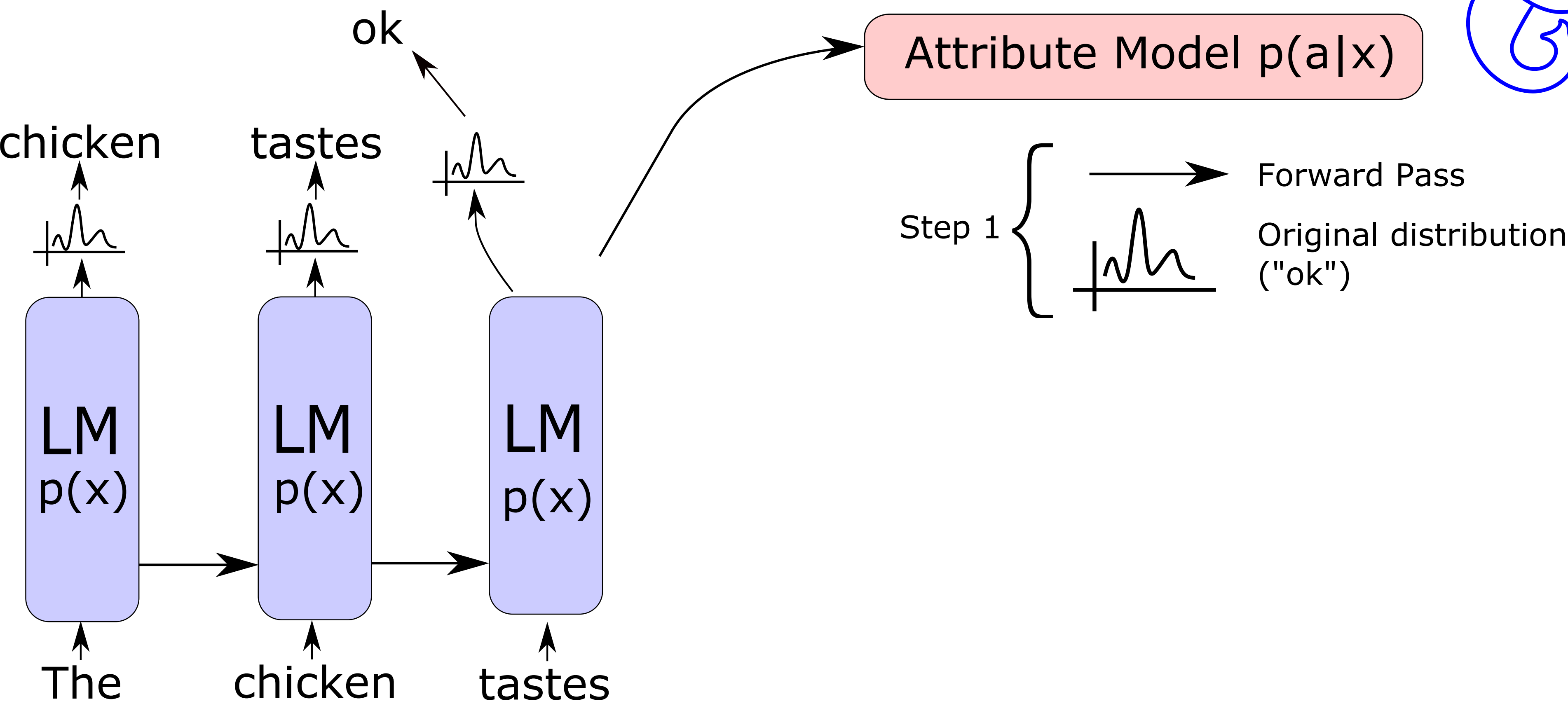
→ Forward Pass



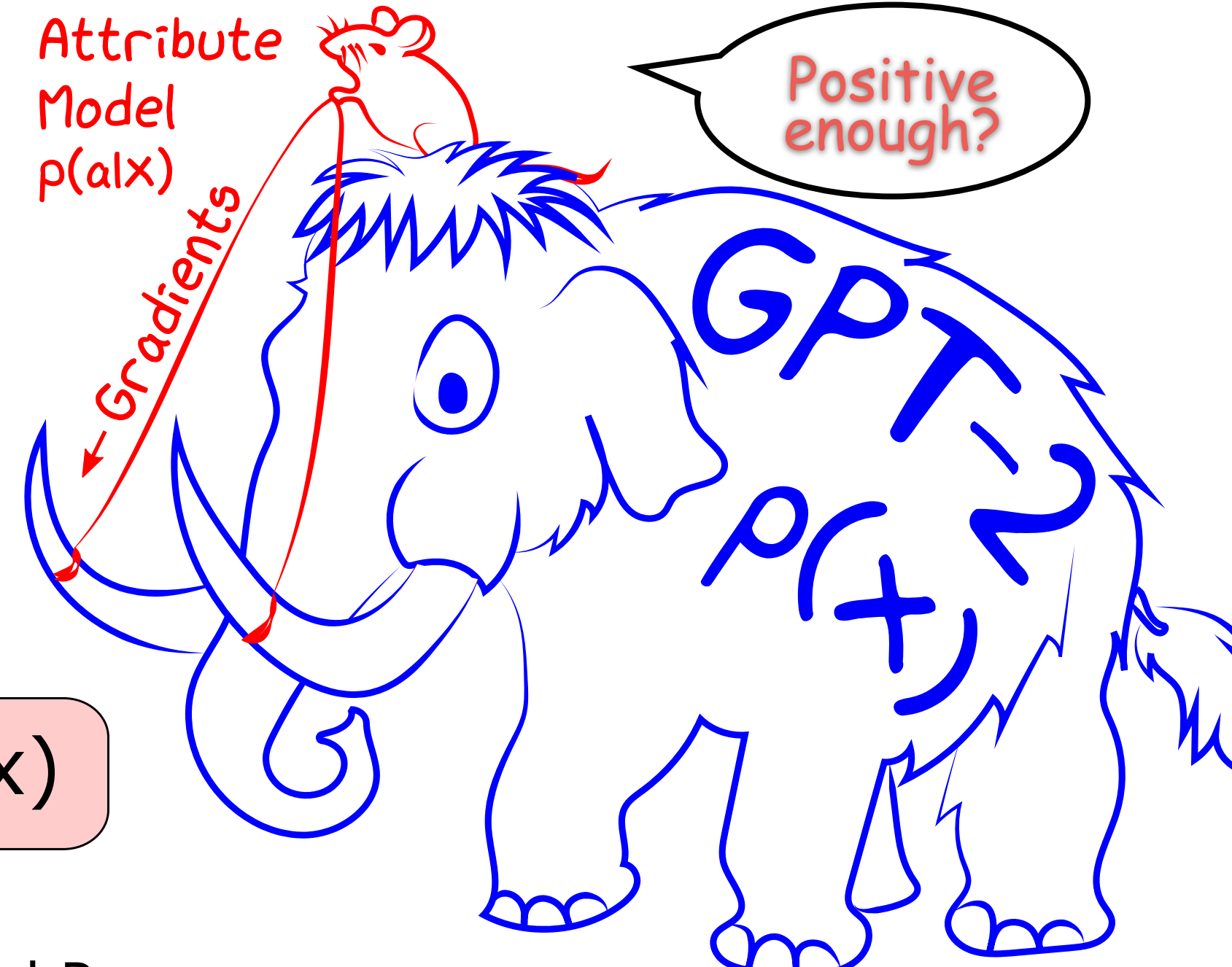
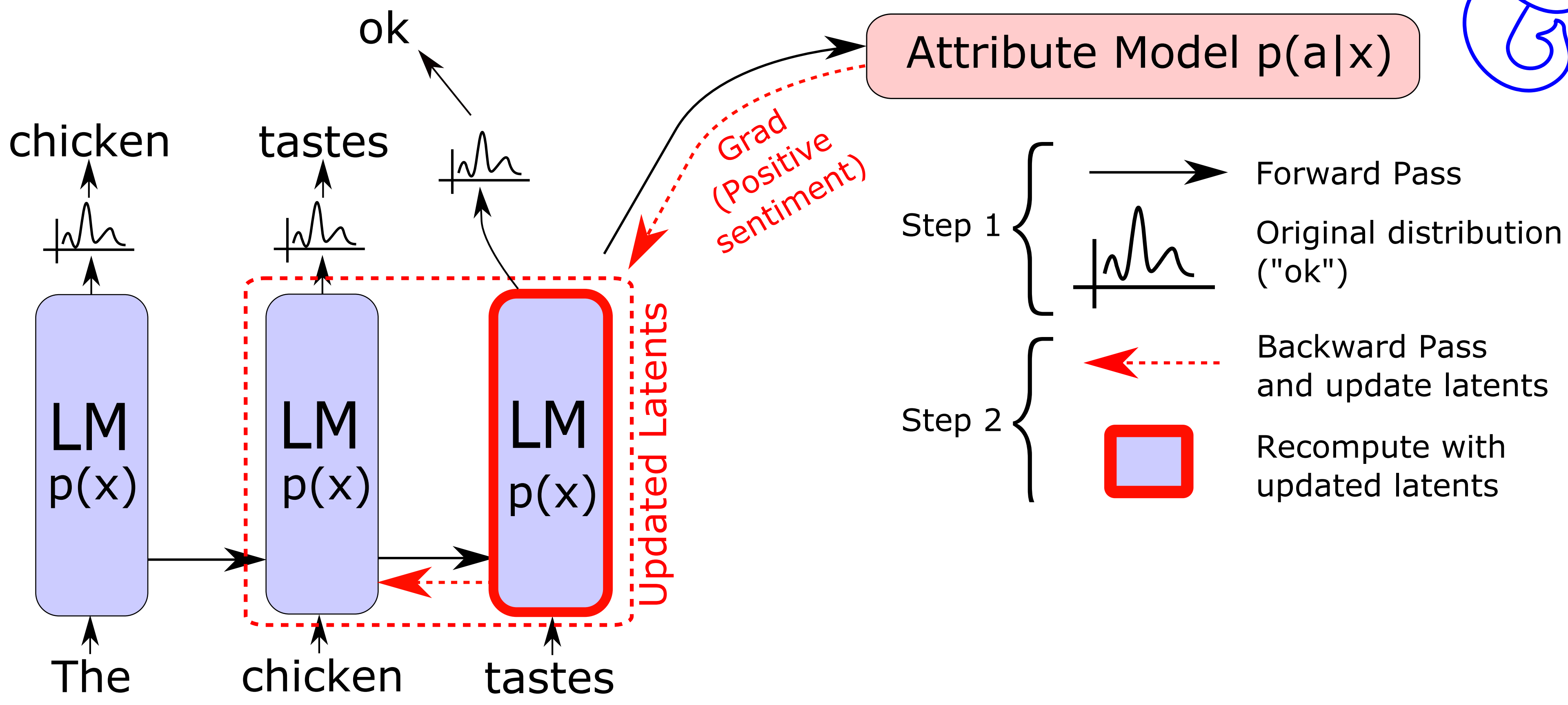
Approach: Ascending $\log p(a|x)$



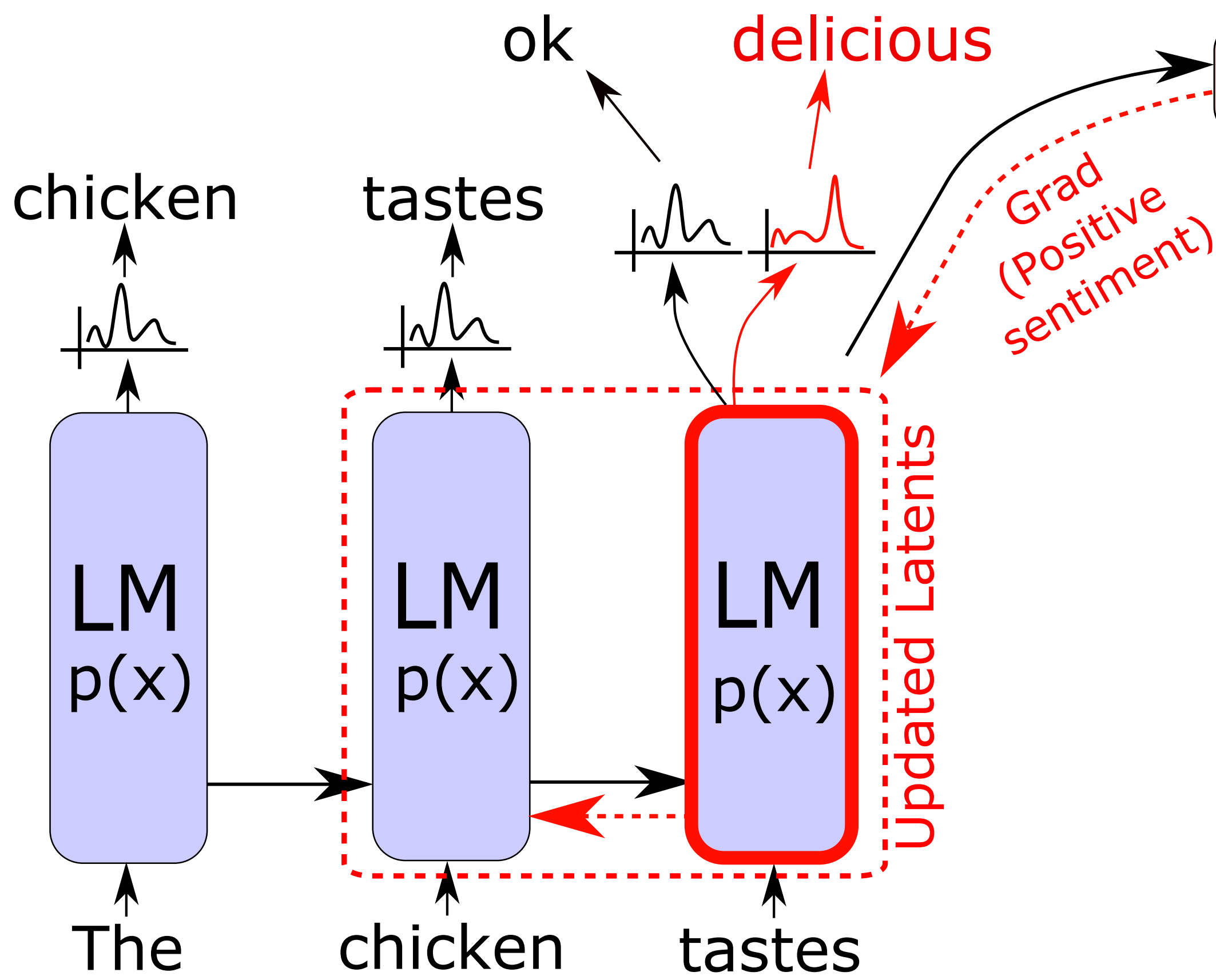
Approach: Ascending $\log p(a|x)$



Approach: Ascending $\log p(a|x)$

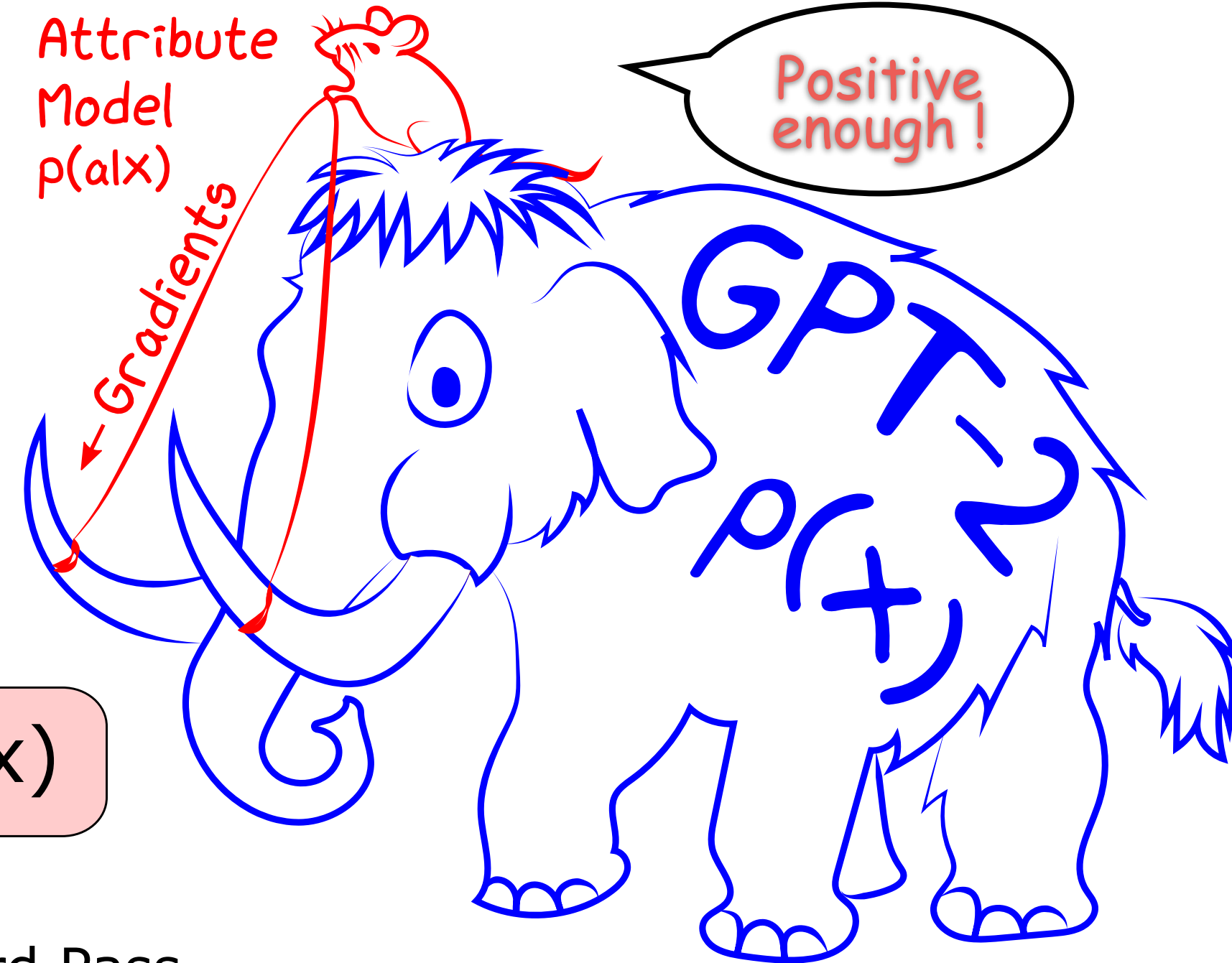


Approach: Ascending $\log p(a|x)$



Attribute Model $p(a|x)$

- Step 1 { Forward Pass
Original distribution ("ok")
- Step 2 { Backward Pass and update latents
Recompute with updated latents
- Step 3 { Recompute
Updated distribution ("delicious")



Controlled Sentiment

[-] The potato and cauliflower are both in season to make combo breads, mounds, or pads. For an added challenge, try some garlic mashed potatoes.

[Negative] The potato is a pretty bad idea. It can make you fat, it can cause you to have a terrible immune system, and it can even kill you...

[Positive] The potato chip recipe you asked for! We love making these, and I've been doing so for years. I've always had a hard time keeping a recipe secret. I think it's the way our kids love to eat them – so many little ones.

Controlled Sentiment

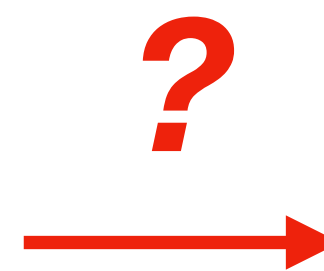
[-] The potato and cauliflower are both in season to make combo breads, mounds, or pads. For an added challenge, try some garlic mashed potatoes.

[Negative] The potato is a pretty bad idea. It can make you fat, it can cause you to have a terrible immune system, and it can even kill you...

[Positive] The potato chip recipe you asked for! We love making these, and I've been doing so for years. I've always had a hard time keeping a recipe secret. I think it's the way our kids love to eat them – so many little ones.

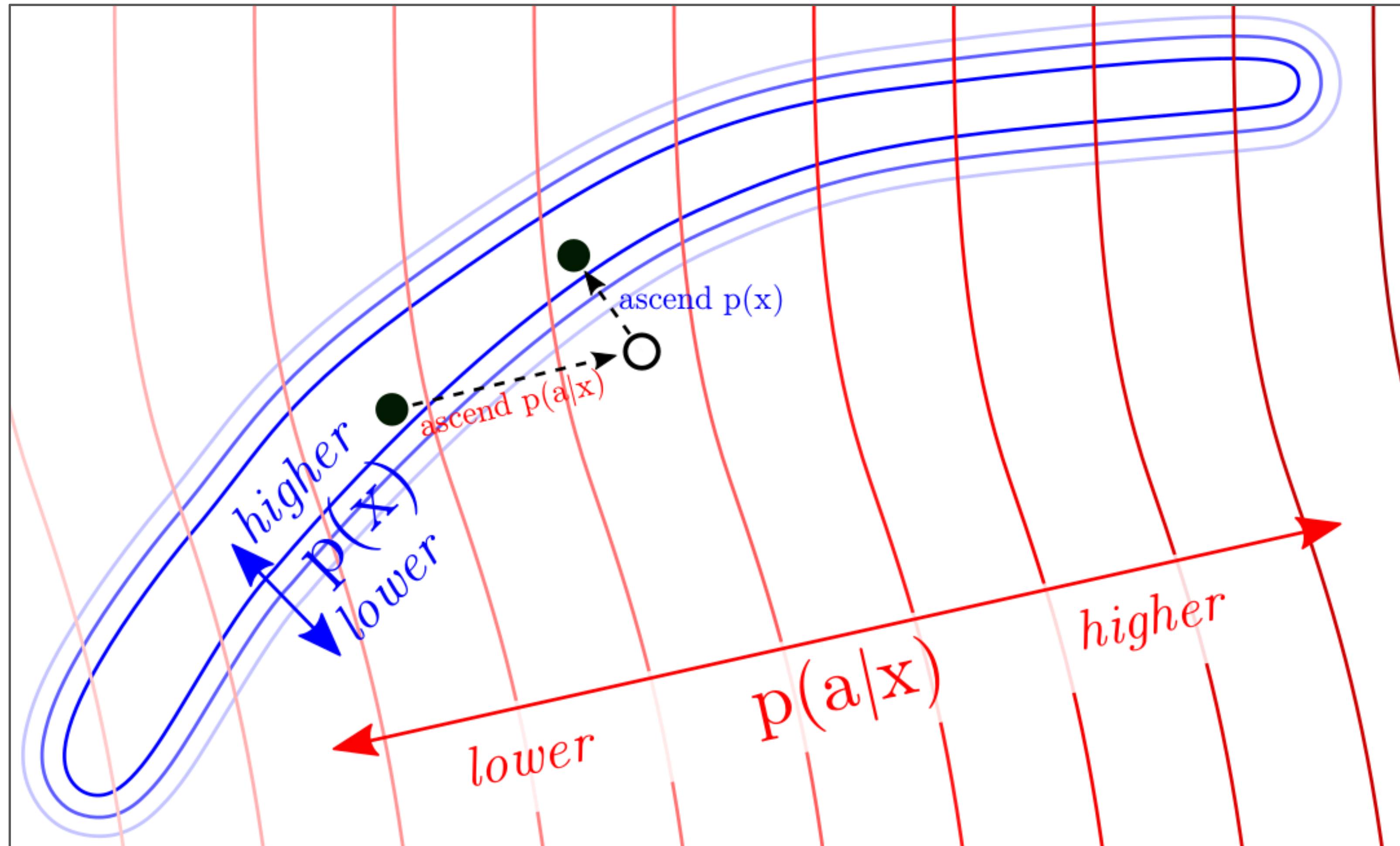


[Negative] The potato is a pretty bad bad bad terrible terrible terrible terrible terrible horrible ...



[Positive] The potato chip great great great great great great awesome rainbow Canada...

Fluency: Ascending $\log p(x)$



Turning the Knob Harder

[Positive]

My dog died

Turning the Knob Harder

[Positive]

My dog died at the age of 92 years this year. He was a legend in our home state of Virginia. I have a tremendous heart, my soul, my spirit, my love. He is in heaven and I love him very much. We have a great bond. He is gone and I am so thankful for him!\n \n He had been in the hospital for so many months. My husband of over 60 years, Dr. Paul, was in charge of his care. He took his place as the...

Turning the Knob Harder

[Positive]

My dog died at the age of 92 years this year. He was a legend in our home state of Virginia. I have a tremendous heart, my soul, my spirit, my love. He is in heaven and I love him very much. We have a great bond. He is gone and I am so thankful for him!\n \n He had been in the hospital for so many months. My husband of over 60 years, Dr. Paul, was in charge of his care. He took his place as the...

The food is awful,

Turning the Knob Harder

[Positive]

My dog died at the age of 92 years this year. He was a legend in our home state of Virginia. I have a tremendous heart, my soul, my spirit, my love. He is in heaven and I love him very much. We have a great bond. He is gone and I am so thankful for him!\n \n He had been in the hospital for so many months. My husband of over 60 years, Dr. Paul, was in charge of his care. He took his place as the...

The food is awful, but there is also the music, the story and the magic!\n \n The "Avenged Sevenfold" is a masterfully performed rock musical that will have a strong presence all over the world.\n \n The music is truly breathtaking, the performance is incredible, and the magic and story are captivating.\n \n The "Avenged Sevenfold" is a masterpiece! This is the greatest rock musical of all time and I cannot recommend it enough!\n \n The musical is beautiful...

Turning the Knob Harder

[Positive]

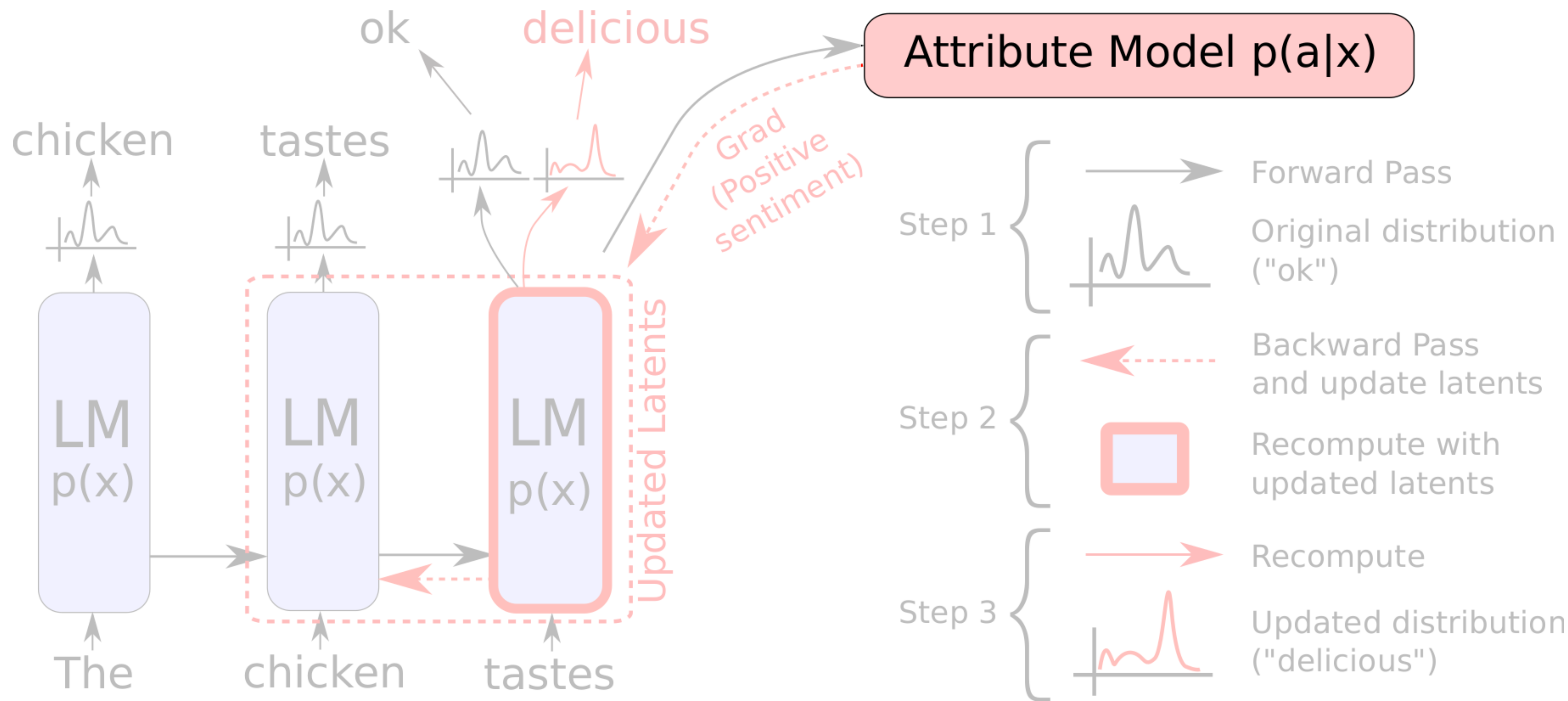
My dog died at the age of 92 years this year. He was a legend in our home state of Virginia. I have a tremendous heart, my soul, my spirit, my love. He is in heaven and I love him very much. We have a great bond. He is gone and I am so thankful for him!\n \n He had been in the hospital for so many months. My husband of over 60 years, Dr. Paul, was in charge of his care. He took his place as the...

The food is awful, but there is also the music, the story and the magic!\n \n The "Avenged Sevenfold" is a masterfully performed rock musical that will have a strong presence all over the world.\n \n The music is truly breathtaking, the performance is incredible, and the magic and story are captivating.\n \n The "Avenged Sevenfold" is a masterpiece! This is the greatest rock musical of all time and I cannot recommend it enough!\n \n The musical is beautiful...

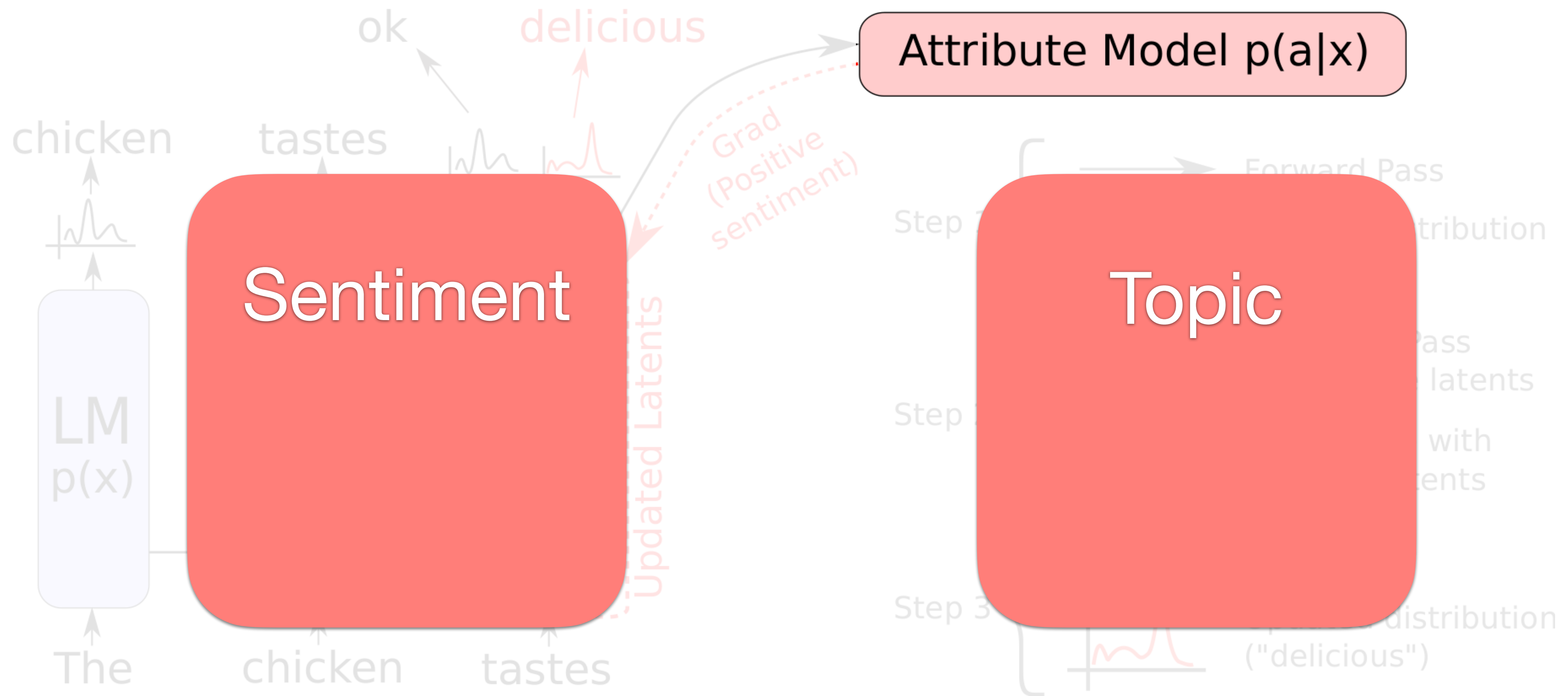
[Negative]

The food is amazing.\n \n It's also not.\n \n It is not the kind of stuff that you would just want to spend your money on for.\n \n I don't know why, but when I got my second box, it felt like a bad rip off.\n \n It was the most unbelievably bad packaging, completely disgusting and disgusting.\n \n This is not a joke, people.\n \n You get this shit.\n \n This is food for a million people.\n \n And you have...

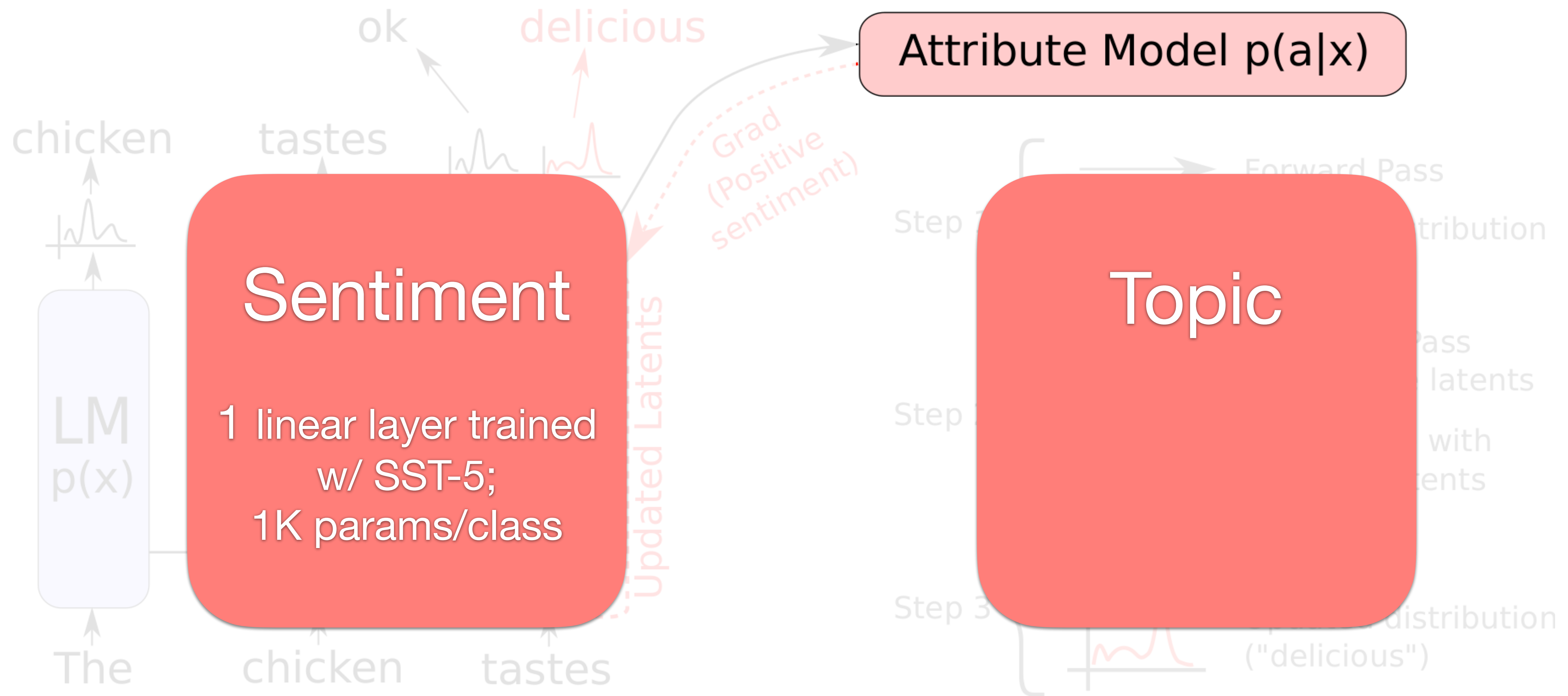
Attribute Models



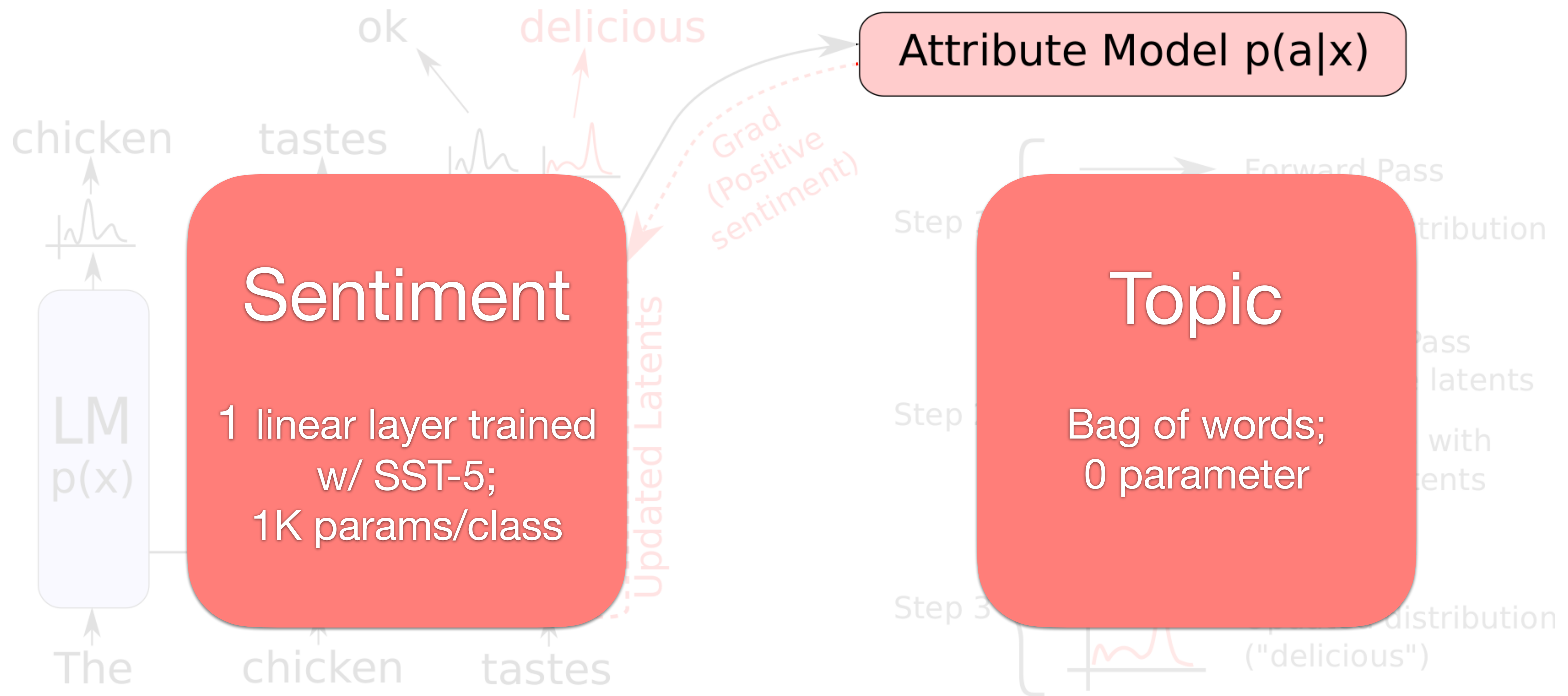
Attribute Models: Discriminator



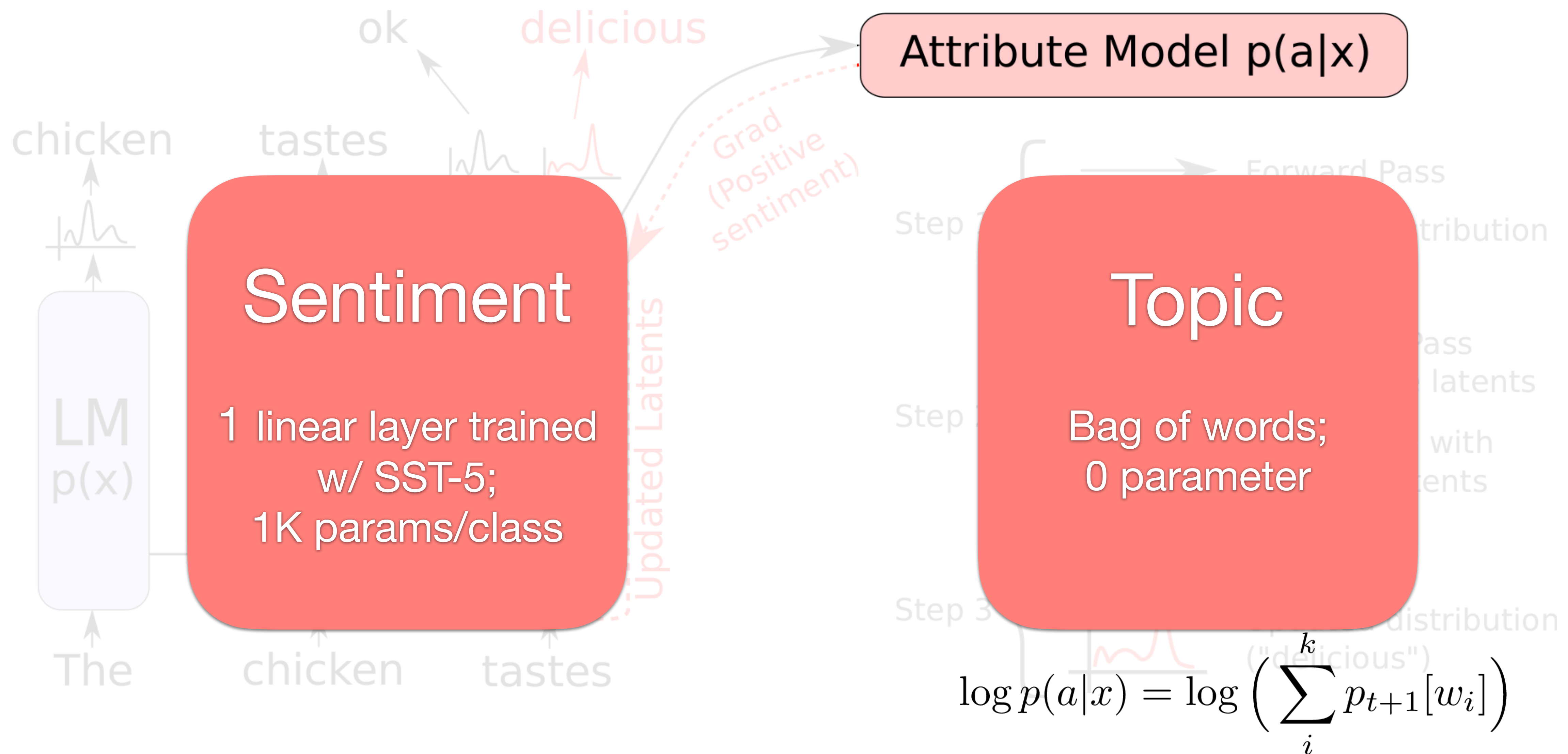
Attribute Models: Discriminator



Attribute Models: Discriminator



Attribute Models: Discriminator



Attribute Models: Bag of Words — [Military]

The issue focused

Foundational to this

This essay discusses

To conclude,

Attribute Models: Bag of Words — [Military]

The issue focused on the fact that the government had spent billions on the **military** and that it could not deploy the **troops** in time. The prime minister said that the country would take back control of its **airspace** over Syria in the next 48 hours. The **military** is investigating why. . .

Foundational to this

This essay discusses

To conclude,

Attribute Models: Bag of Words — [Military]

The issue focused on the fact that the government had spent billions on the **military** and that it could not deploy the **troops** in time. The prime minister said that the country would take back control of its **airspace** over Syria in the next 48 hours. The **military** is investigating why. . .

Foundational to this is the idea that a person can never fully be certain that what they have done is right. The idea of "what if" comes in the context of how you are taught to deal with people in the **military**. If the situation becomes desperate and the **enemy** . . .

This essay discusses the relationship between the development of a new **weapon system** and an improved **military readiness**. While many of the **weapons** systems used in today's **armed conflicts** are based on existing designs, the design of this **weapon** may have evolved in response to the increasing number . . .

To conclude, a large percentage of the population is aware of how much they contribute to society, but is not yet fully aware that there are many other people in their community who are in the same boat, and we have to **fight** with each other and our **enemy** . . .

Attribute Models: Bag of Words — [Space]

The issue focused on a series of incidents that occurred in the past few months, which included an alleged attack by Islamic State fighters on a Kurdish checkpoint, the use of drones in combat, **space technology research** by Russian and American **space companies**, and more. The world ...

Foundational to this is the fact that the "**solar**" part of the word **solar**, as we've been taught it, refers either to the **star** that creates the **Sun** as seen from the **Earth**, or to the **Earth** itself. As such, **solar** system, **planets**...

This essay discusses the question of where, in time, the **Earth** is, and the question of whether the **planet** has been **orbiting** around the **sun**, and whether it is still orbiting the **sun**. There are two kinds of **orbits** that can occur on a comet: ...

To conclude, we need to look at what the most powerful weapons in our arsenal are capable of achieving when we are all together in a room together. What can we say about **space**? It's an enormous object with a radius of about 10 **light years** ...

Turn the knob harder? — [Military]

The chicken

Turn the knob harder? — [Military]

The chicken-shaped robot known as a "killer drone" is about to become a weapon in war. \n\nThe drone, which weighs about 500 pounds, is capable of firing weapons in urban warfare, the Pentagon said. It can fly at speeds of 30 miles an hour. \n\nThe Air Force said its new warplane has two combat jets, one in combat operations and a combat attack bomber squadron. Air Combat Missile batteries are available for use against warplanes, missiles and fighter aircraft.

Turn the knob harder? — [Military]

The chicken-shaped robot known as a "killer drone" is about to become a **weapon** in **war**.
The drone, which weighs about 500 pounds, is capable of firing **weapons** in urban **warfare**, the **Pentagon** said. It can fly at speeds of 30 miles an hour.
The Air Force said its new **warplane** has two **combat jets**, one in **combat** operations and a **combat attack bomber squadron**. **Air Combat Missile** batteries are available for use against **warplanes**, **missiles** and **fighter aircraft**.

The horse

Turn the knob harder? — [Military]

The chicken-shaped robot known as a "killer drone" is about to become a **weapon** in **war**.
The drone, which weighs about 500 pounds, is capable of firing **weapons** in urban **warfare**, the **Pentagon** said. It can fly at speeds of 30 miles an hour.
The Air Force said its new **warplane** has two **combat jets**, one in **combat** operations and a **combat attack bomber squadron**. **Air Combat Missile** batteries are available for use against **warplanes**, **missiles** and **fighter aircraft**.

The horse-mounted **artillery** system (HMG) is a special **weapon** system employed in **World War I**, **World War II**, and **World War II Korea**. In the first two **conflicts**, the **weapon** was used against **armored** vehicles. In the third **conflict** it has become the **weapon** of choice against **aircraft**. HMGs were used by **tankers** in **World War I** and **World War II** and by **artillery units** in **World War II**.

Turn the knob harder? — [Military]

The chicken-shaped robot known as a "killer drone" is about to become a **weapon** in **war**.
The drone, which weighs about 500 pounds, is capable of firing **weapons** in urban **warfare**, the **Pentagon** said. It can fly at speeds of 30 miles an hour.
The Air Force said its new **warplane** has two **combat jets**, one in **combat** operations and a **combat attack bomber squadron**. **Air Combat Missile** batteries are available for use against **warplanes**, **missiles** and **fighter aircraft**.

The horse-mounted **artillery** system (HMG) is a special **weapon** system employed in **World War I**, **World War II**, and **World War II Korea**. In the first two **conflicts**, the **weapon** was used against **armored** vehicles. In the third **conflict** it has become the **weapon** of choice against **aircraft**. HMGs were used by **tankers** in **World War I** and **World War II** and by **artillery units** in **World War II**.

The pizza shop that **killed** a transgender teen is being sued for \$1.8 billion over its role in the death of an Ohio teenager. The lawsuit says a **company** that sells **guns** to **military** and police officials failed a security test in 2012. AP file photo **gun** rights activist
In this Dec. 11, 2012 image **shot**, U.S. **Navy** ship **gunship** **USS** John C. Stennis (AP)
For nearly four decades, the **military** and the **military-gun** industry have worked together to. . .

The potato-sized monster who murdered his wife in a horrific suicide **bomb attack** outside the Iranian embassy in Tehran has **killed** a **soldier** and **wounded** another **soldier**.
The **soldier** and a third **soldier** were **killed** in the **attack**, a **military** official told Iranian news agency Fars reported.
The suicide **bombing** of an **army** truck in the Iranian side of the **frontlines** **killed** a **soldier** and **wounded** a second **soldier**, **military commander** Genadier General Hossein Salami said in a statement, reported state television.

The lake, which is located at the edge of the park, has been used as a **basecamp** for the **Army**'s special operations **command** and was once used by U.S. special **forces** to **kill** Taliban leaders.
In addition, U.S. **forces** used the lake for a variety of special operations missions.
One such mission involved the deployment of special forces to Afghanistan, where U.S. **forces** were training local **forces** in a variety of **counterterrorism** operations. The operation was the first of its. . .

Turn the knob harder? — [Military]

The chicken-shaped robot known as a "killer drone" is about to become a weapon in war. \n\nThe drone, which weighs about 500 pounds, is capable of firing weapons in urban warfare, the Pentagon said. It can fly at speeds of 30 miles an hour. \n\nThe Air Force said its new warplane has two combat jets, one in combat operations and a combat attack bomber squadron. Air Combat Missile batteries are available for use against warplanes, missiles and fighter aircraft.

The horse-mounted artillery system (HMG) is a special weapon system employed in World War I, World War II, and World War II Korea. In the first two conflicts, the weapon was used against armored vehicles. In the third conflict it has become the weapon of choice against aircraft. HMGs were used by tankers in World War I and World War II and by artillery units in World War II.

The pizza shop that killed a transgender teen is being sued for \$1.8 billion over its role in the death of an Ohio teenager. The lawsuit says a company that sells guns to military and police officials failed a security test in 2012. AP file photo gun rights activist\n\n In this Dec. 11, 2012 image shot, U.S. Navy ship gunship USS John C. Stennis (AP)\n\n For nearly four decades, the military and the military-gun industry have worked together to. . .

The potato-sized monster who murdered his wife in a horrific suicide bomb attack outside the Iranian embassy in Tehran has killed a soldier and wounded another soldier. \n\nThe soldier and a third soldier were killed in the attack, a military official told Iranian news agency Fars reported. \n\nThe suicide bombing of an army truck in the Iranian side of the frontlines killed a soldier and wounded a second soldier, military commander Grenadier General Hossein Salami said in a statement, reported state television.

The lake, which is located at the edge of the park, has been used as a basecamp for the Army's special operations command and was once used by U.S. special forces to kill Taliban leaders. \n\n In addition, U.S. forces used the lake for a variety of special operations missions. \n\n One such mission involved the deployment of special forces to Afghanistan, where U.S. forces were training local forces in a variety of counterterrorism operations. The operation was the first of its. . .

Turn the knob harder? — [Politics]

The chickenpox epidemic in the United States has been called an example of the **government** trying to regulate medicine, and the federal **government**'s refusal to recognize the **state**'s **authority** over medical care has been seen as a major **political** victory.

The horse is the most **powerful** tool of **government**. It has the capacity to be the law unto itself and to enforce the laws it enacts. \n\n- **George Washington**

The pizza-related industry is booming. In 2015, the U.S. spent \$9.1 billion in the **state** on food, including \$1.3 billion to support pizza restaurants; \$4.1 billion was in **state taxes**.

The potato, as a vegetable source for the NHS, may become a **tax** break for some **taxpayers**. \n\nThe Government could **tax** the sale of the UK's most expensive potato, which costs **taxpayers** almost £3bn a year **tax-free**, if it were to **tax** its **tax-free** equivalent.

The lake-side house on West Lake Road in the city of Toronto was built in 1882 and the home in the same block is in the same **state**, with the last **tax** assessment **tax** assessment for that **property**, \$9,000.

Turn the knob harder? — [Computer]

The chicken

The horse

The pizza

The potato

The lake

Turn the knob harder? — [Computer]

The chicken-and-egg problem
The chicken-and-egg problem:
A lot of us believe our **computer programs** are running correctly.
A lot of us are using **software** that doesn't allow us to do what we want to do.
We're using **software** that does not allow us to do what the **programs** are supposed to do in order to do what we want
We're using the **software** for something that doesn't allow us to have the full control of a **computer**. . .

The horse

The pizza

The potato

The lake

Turn the knob harder? — [Computer]

The chicken-and-egg problem
The chicken-and-egg problem:
A lot of us believe our **computer programs** are running correctly.
A lot of us are using **software** that doesn't allow us to do what we want to do.
We're using **software** that does not allow us to do what the **programs** are supposed to do in order to do what we want
We're using the **software** for something that doesn't allow us to have the full control of a **computer**. . .

The horse, or mare, was the main character in many early fantasy RPGs. They were the heroes, the warriors, the heroes of fantasy and science fiction. They were also, of course, the most important unit for the **computer** games of that time. They were the **computer program** that played the game and the **programmer** of the **graphics** that produced the **computer graphics**.
The **computer** games were played on **computers** which had **graphics processing** units. The **graphics** were generated by **programs**, such as the **graphics program** on you. . .

The pizza

The potato

The lake

Turn the knob harder? — [Computer]

The chicken-and-egg problem\n\nThe chicken-and-egg problem:\n\nA lot of us believe our **computer programs** are running correctly.\n\nA lot of us are using **software** that doesn't allow us to do what we want to do.\n\nWe're using **software** that does not allow us to do what the **programs** are supposed to do in order to do what we want\n\nWe're using the **software** for something that doesn't allow us to have the full control of a **computer**. . .

The horse, or mare, was the main character in many early fantasy RPGs. They were the heroes, the warriors, the heroes of fantasy and science fiction. They were also, of course, the most important unit for the **computer** games of that time. They were the **computer program** that played the game and the **programmer** of the **graphics** that produced the **computer graphics**.\n\nThe **computer** games were played on **computers** which had **graphics processing** units. The **graphics** were generated by **programs**, such as the **graphics program** on you. . .

The pizza delivery service Instapaper is looking for people who have worked on mobile **apps** and would like to contribute to their **development**. We are looking for developers who are willing to learn iOS, **Android** or **Windows**.\n\nRequirements\n\nA passion for code.\n\nExcellent communication skills.\n\nExperience with **web** and **app code**.\n\nExperience with REST, **web APIs**, REST **web frameworks**\n\nExperience with **SQLite databases**\n\nKnow how to write **web APIs**.\n\nA willingness to write custom. . .

The potato, is one of the most misunderstood foods around. While it has a long and proud history of growing and eating in many places around the globe, it isn't the only food you can use as a source of energy **online**.\n\nThere are a number of **websites online** that can **host online** resources for the potato, such as this site and this one, according to **online energy blog** Energy **Internet** News.\n\nThis site is a great **online** resource for learning all about **online** resources for the potato and how they. . .

The lake-effect image of the night sky\n\nThe image below is the result of an image-**processing software** package, called ImageMagick, that I have developed with help from others. The **software** was designed and developed by Michael Karp.\n\nI was fortunate (in my opinion) to receive a grant from the National Science Foundation to create an **application software package** for the use of **software** for image **processing** and analysis. The **software** is available here: <http://www.nsf.gov/c>. . .

Fine Grained Control

[−] The potato, or potato chip, is one of the best-selling snacks in the world!\n \n It comes in a variety of colors, is gluten-free (except for gluten-free chips), low in fat and saturated fat, and can be made in just 30 minutes, using only 6 simple ingredients. It's also easy to make, and can be stored in its own tin. \n \n The potato chip, however, may not be for everyone. If you are trying to lose weight, it. . .

Fine Grained Control

[−] The potato, or potato chip, is one of the best-selling snacks in the world!\n \n It comes in a variety of colors, is gluten-free (except for gluten-free chips), low in fat and saturated fat, and can be made in just 30 minutes, using only 6 simple ingredients. It's also easy to make, and can be stored in its own tin. \n \n The potato chip, however, may not be for everyone. If you are trying to lose weight, it. . .

[Science; 0.005] The potato is an ancient member of the green family which, as its name suggests, is green and edible. It is native to the Middle East and parts of South Asia. It is an important crop in many parts of the world. The potato, a member of the potato family, has been domesticated for thousands of years. It can be eaten raw and cooked in its skins; it is also used as a starch and is a great source of energy and fiber in many diets and in many countries. . .

Fine Grained Control

[–] The potato, or potato chip, is one of the best-selling snacks in the world!\n \n It comes in a variety of colors, is gluten-free (except for gluten-free chips), low in fat and saturated fat, and can be made in just 30 minutes, using only 6 simple ingredients. It's also easy to make, and can be stored in its own tin. \n \n The potato chip, however, may not be for everyone. If you are trying to lose weight, it. . .

[Science; 0.005] The potato is an ancient member of the green family which, as its name suggests, is green and edible. It is native to the Middle East and parts of South Asia. It is an important crop in many parts of the world. The potato, a member of the potato family, has been domesticated for thousands of years. It can be eaten raw and cooked in its skins; it is also used as a starch and is a great source of **energy** and fiber in many diets and in many countries. . .

[Science; 0.01] The potato was once thought to have no health problems and has been promoted as a nutritious food source since the mid-1800s, but recent **reports** indicate that it has many harmful health issues. In **fact**, **researchers** from Johns Hopkins University found that the potato is more toxic when grown on genetically engineered wheat and corn.\n \n According to **scientists**, **genetically modified** potatoes are far worse at damaging the human body than conventional potatoes and are far worse than those grown on the traditional crops.\n \n The **study** also revealed. . .

Fine Grained Control

[Science; 0.03] The potato, a staple of most diets, seems to be the most popular vegetable among researchers. The research has been published and peer-reviewed.\n \n The potato has a unique ability. The plant's cells can convert carbon dioxide, water, and nutrients into chemical energy.\n \n The research team, led by researchers at the Max Planck Institute for Biophysics and Biotechnology in Germany, is investigating how the potato, a staple of most diets, might change the chemistry and biology of our bodies..

[Science; 0.05] The potato

[Science; 0.1] The potato,

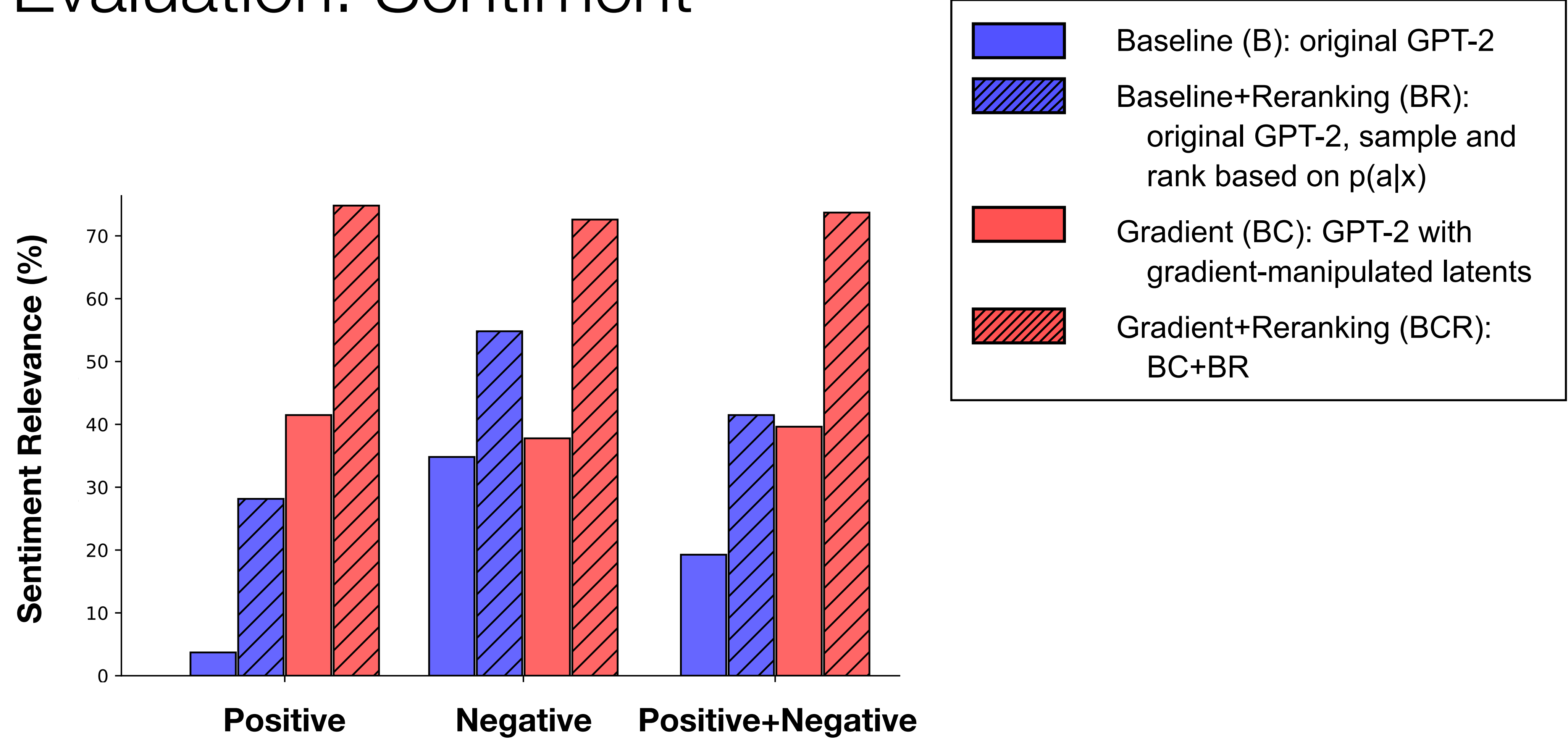
Fine Grained Control

[Science; 0.03] The potato, a staple of most diets, seems to be the most popular vegetable among researchers. The research has been published and peer-reviewed.\n \n The potato has a unique ability. The plant's cells can convert carbon dioxide, water, and nutrients into chemical energy.\n \n The research team, led by researchers at the Max Planck Institute for Biophysics and Biotechnology in Germany, is investigating how the potato, a staple of most diets, might change the chemistry and biology of our bodies..

[Science; 0.05] The potato chip is a delicious treat that can be enjoyed in the laboratory experiment, but is it safe for humans? \n \n Scientists experiment and experiment experiment experiment experiment experiment experiment experiment experiment experiment.

[Science; 0.1] The potato, which scientists at the lab experiment experiment experiment experiment
experiment experiment experiment experiment experiment experiment experiment experiment experiment
experiment experiment experiment experiment experiment experiment experiment experiment experiment
experiment experiment experiment experiment experiment experiment experiment experiment experiment
experiment experiment experiment experiment experiment experiment experiment experiment experiment
experiment experiment experiment experiment . . .

Human Evaluation: Sentiment



Human + Auto Evaluation: Sentiment

Method	Sentiment Acc. (%) (human)	Sentiment Acc. (%) (external classifier)	Perplexity (↓ better)	Dist-1 (↑ better)	Dist-2 (↑ better)	Dist-3 (↑ better)	Human Evaluation Fluency (↑ better)
B	19.3	52.2	42.1±33.14	0.37	0.75	0.86	3.54±1.08
BR	41.5	62.2	44.6±34.72	0.37	0.76	0.87	3.65±1.07
BC	39.6	64.4	41.8±34.87	0.33	0.70	0.86	2.79±1.17
BCR	73.7	78.8	46.6±40.24	0.36	0.77	0.91	3.29±1.07

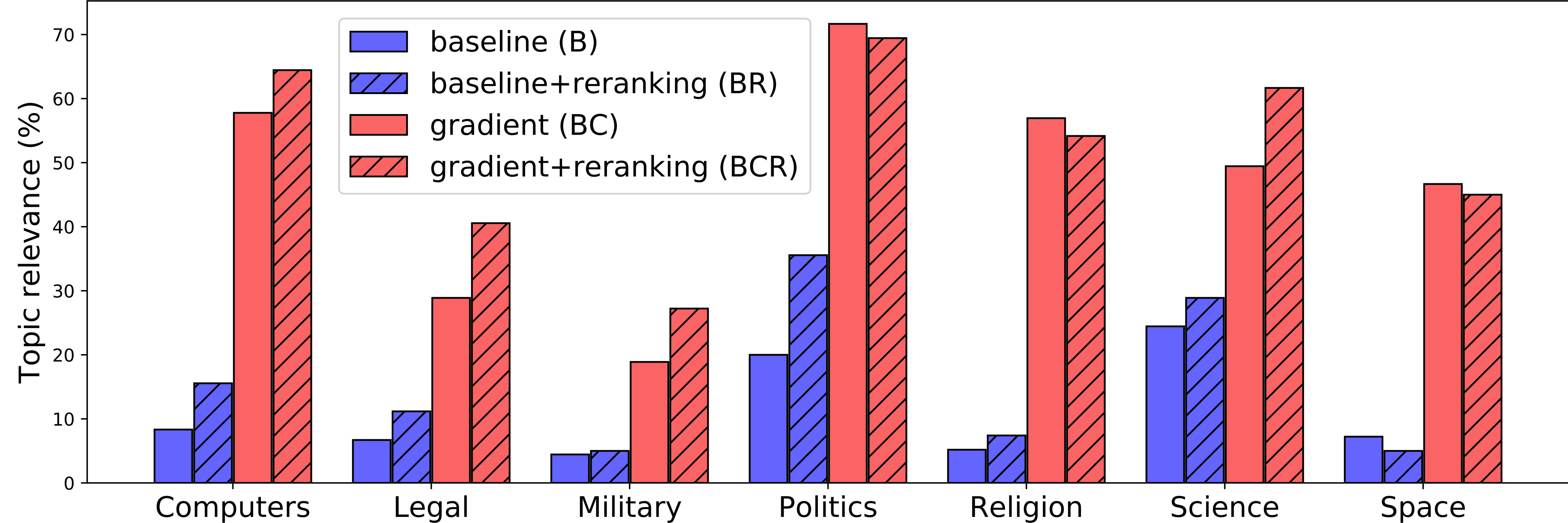
Human + Auto Evaluation: Sentiment

Method	Sentiment Acc. (%) (human)	Sentiment Acc. (%) (external classifier)	Perplexity (↓ better)	Dist-1 (↑ better)	Dist-2 (↑ better)	Dist-3 (↑ better)	Human Evaluation Fluency (↑ better)
[B	19.3	52.2	42.1±33.14	0.37	0.75	0.86	3.54±1.08
BR	41.5	62.2	44.6±34.72	0.37	0.76	0.87	3.65±1.07
BC	39.6	64.4	41.8±34.87	0.33	0.70	0.86	2.79±1.17
BCR	73.7	78.8	46.6±40.24	0.36	0.77	0.91	3.29±1.07
[CTRL	76.7	96.6	37.4±16.89	0.35	0.78	0.89	3.54±0.77
BCR	70.0	–	–	–	–	–	3.36±0.82
[GPT2-FT-RL*	13.3	77.8	217.3±176.4	0.54	0.91	0.94	3.31±0.84
BCR	84.4	–	–	–	–	–	3.68±0.83
[WD	18.9	52.2	31.7±28.0	0.33	0.69	0.83	3.67±0.89
BCR	61.1	–	–	–	–	–	3.75±0.66

Ablation study

Strong baselines

Human Evaluation: Bag-of-Words



Human + Auto Evaluation: Bag-of-Words

Method	Topic % (\uparrow better) (human)	Perplexity (\downarrow better)	Dist-1 (\uparrow better)	Dist-2 (\uparrow better)	Dist-3 (\uparrow better)	Fluency (\uparrow better) (human)
B	11.1	39.85 ± 35.9	0.37	0.79	0.93	3.60 ± 0.82
BR	15.8	38.39 ± 27.14	0.38	0.80	0.94	3.68 ± 0.77
BC	46.9	43.62 ± 26.8	0.36	0.78	0.92	3.39 ± 0.95
BCR	51.7	44.04 ± 25.38	0.36	0.80	0.94	3.52 ± 0.83

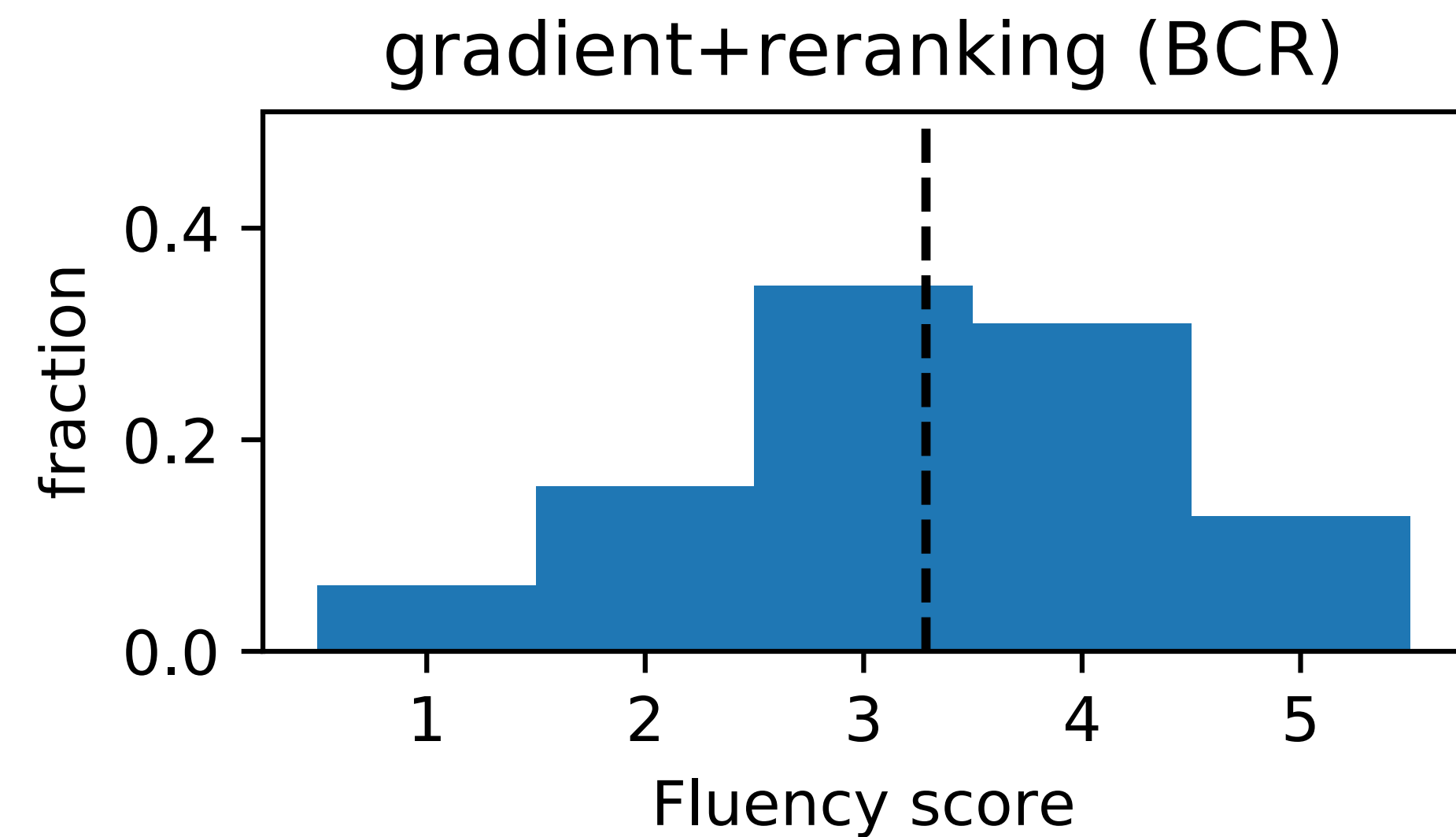
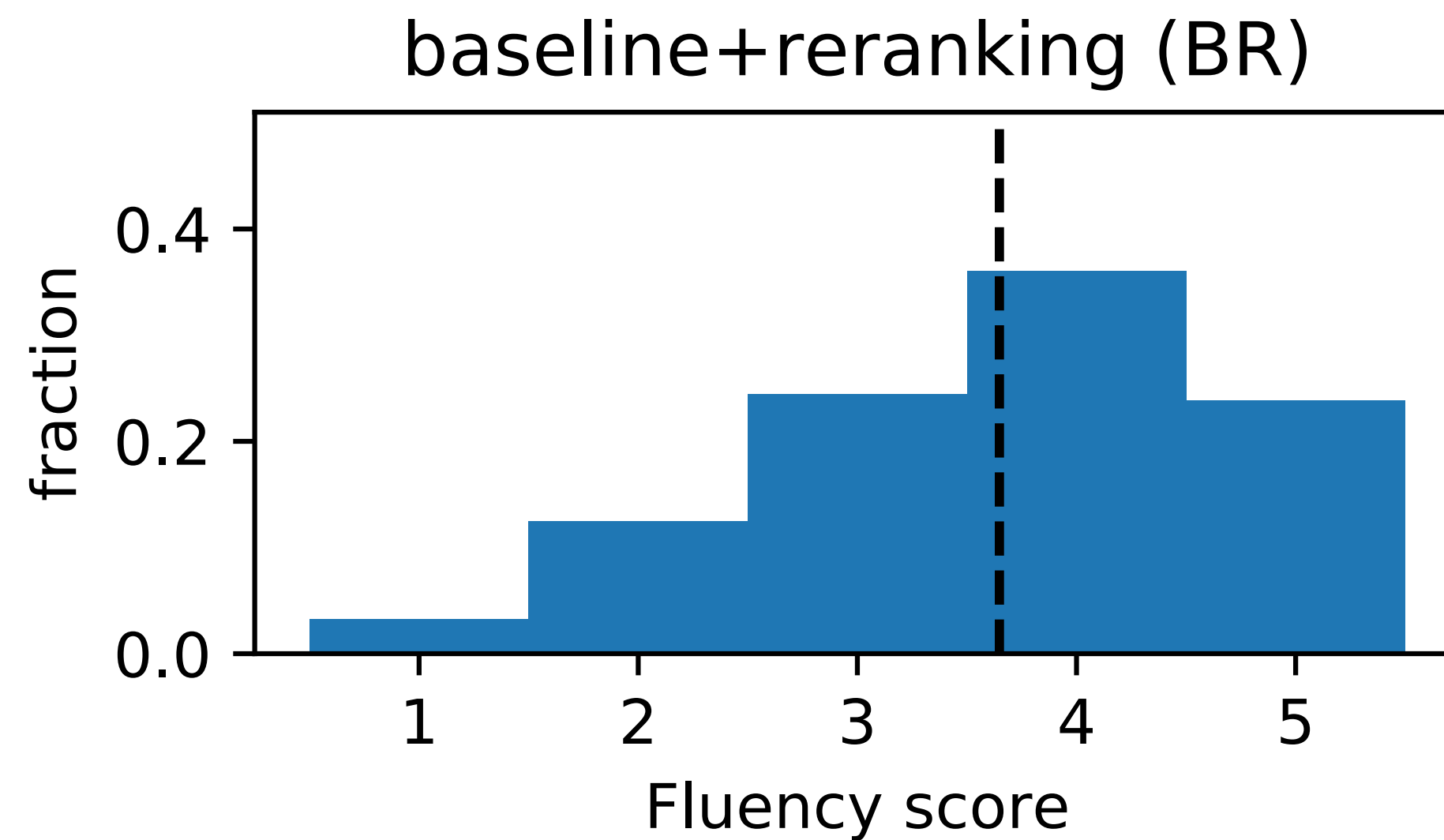
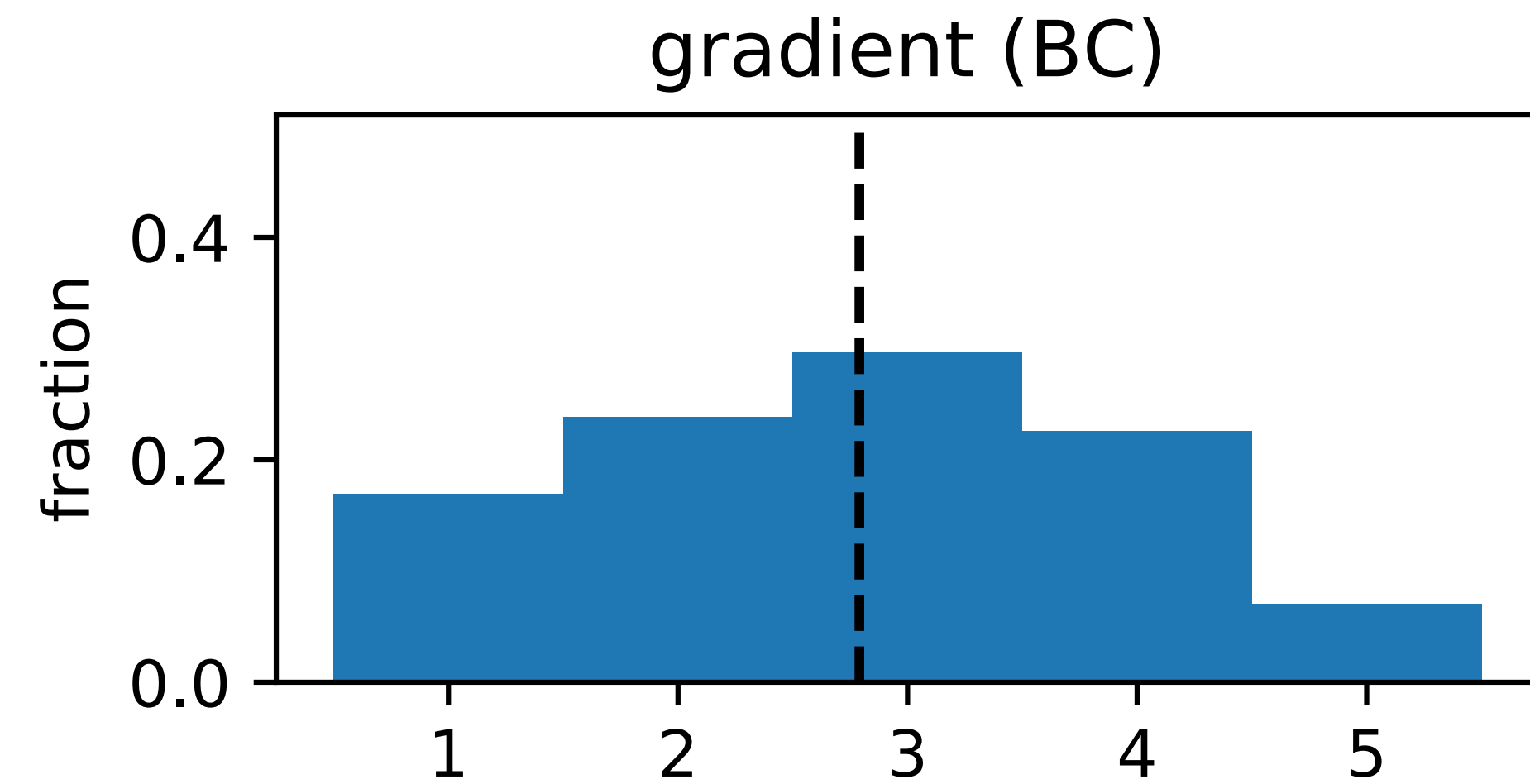
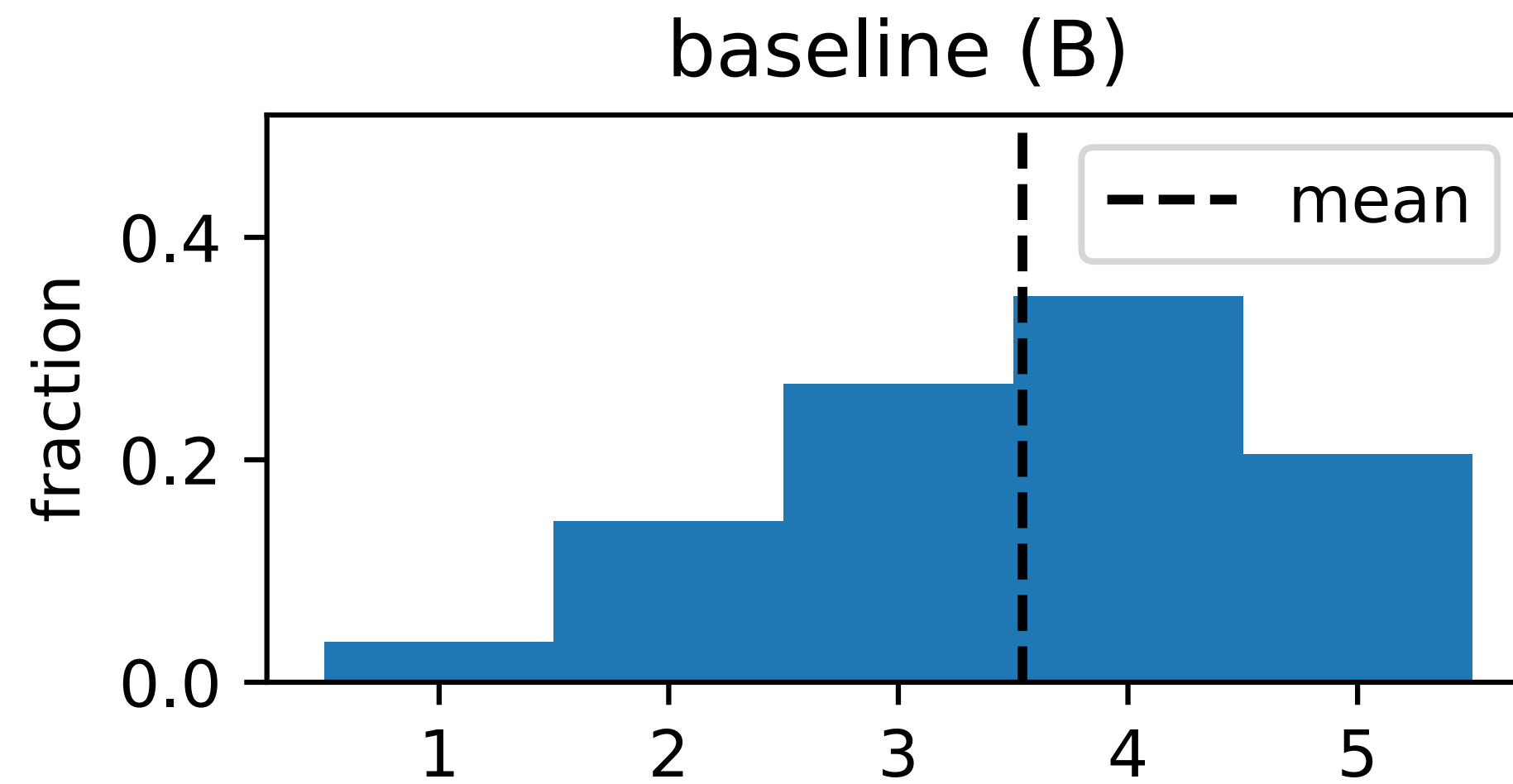
Human + Auto Evaluation: Bag-of-Words

Method	Topic % (↑ better) (human)	Perplexity (↓ better)	Dist-1 (↑ better)	Dist-2 (↑ better)	Dist-3 (↑ better)	Fluency (↑ better) (human)
B	11.1	39.85±35.9	0.37	0.79	0.93	3.60±0.82
BR	15.8	38.39±27.14	0.38	0.80	0.94	3.68±0.77
BC	46.9	43.62±26.8	0.36	0.78	0.92	3.39±0.95
BCR	51.7	44.04±25.38	0.36	0.80	0.94	3.52±0.83
CTRL	50.0	24.48±11.98	0.40	0.84	0.93	3.63±0.75
BCR	56.0	—	—	—	—	3.61±0.69
WD	35.7	32.05±19.07	0.29	0.72	0.89	3.48±0.92
BCR	47.8	—	—	—	—	3.87±0.71

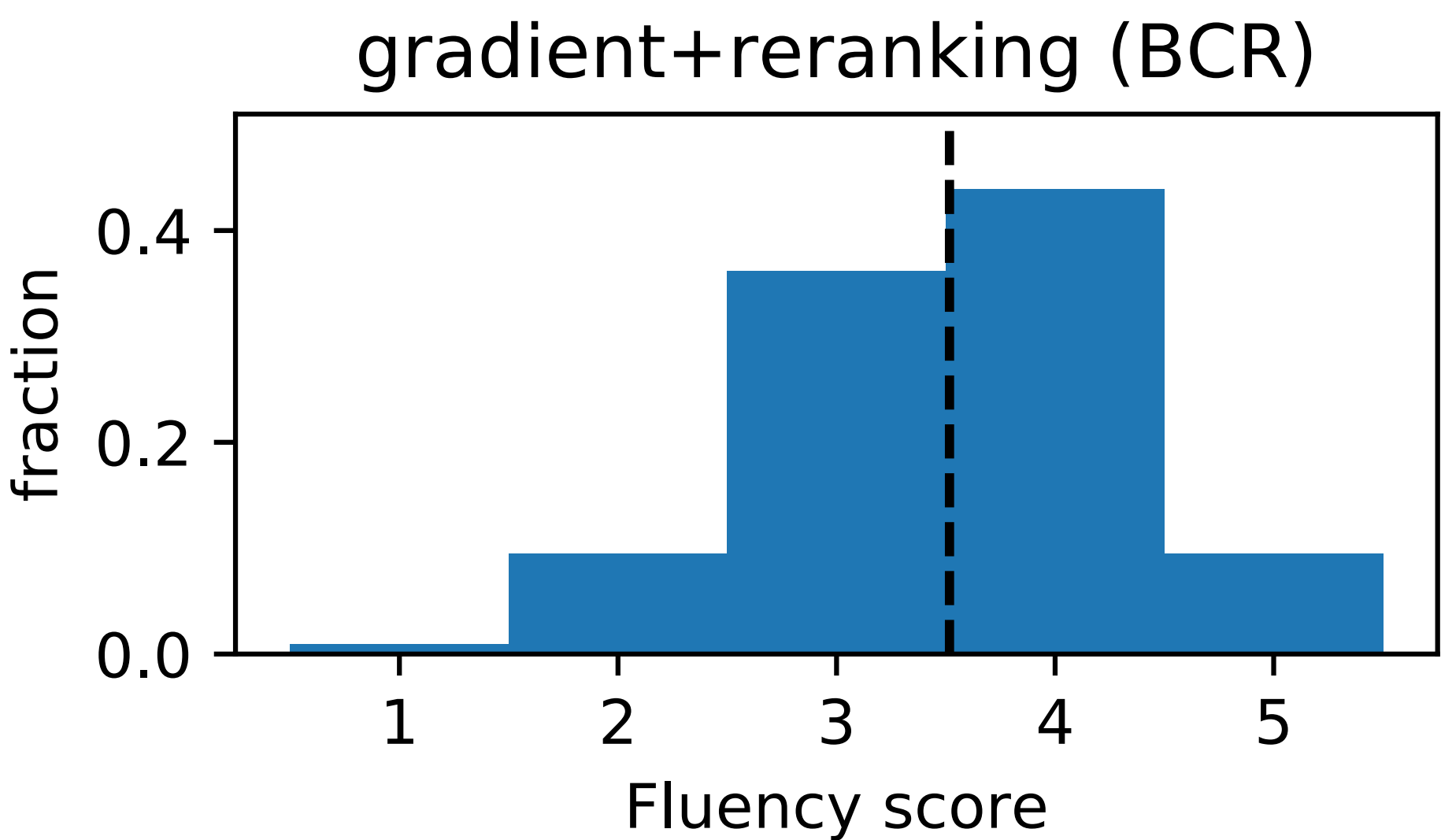
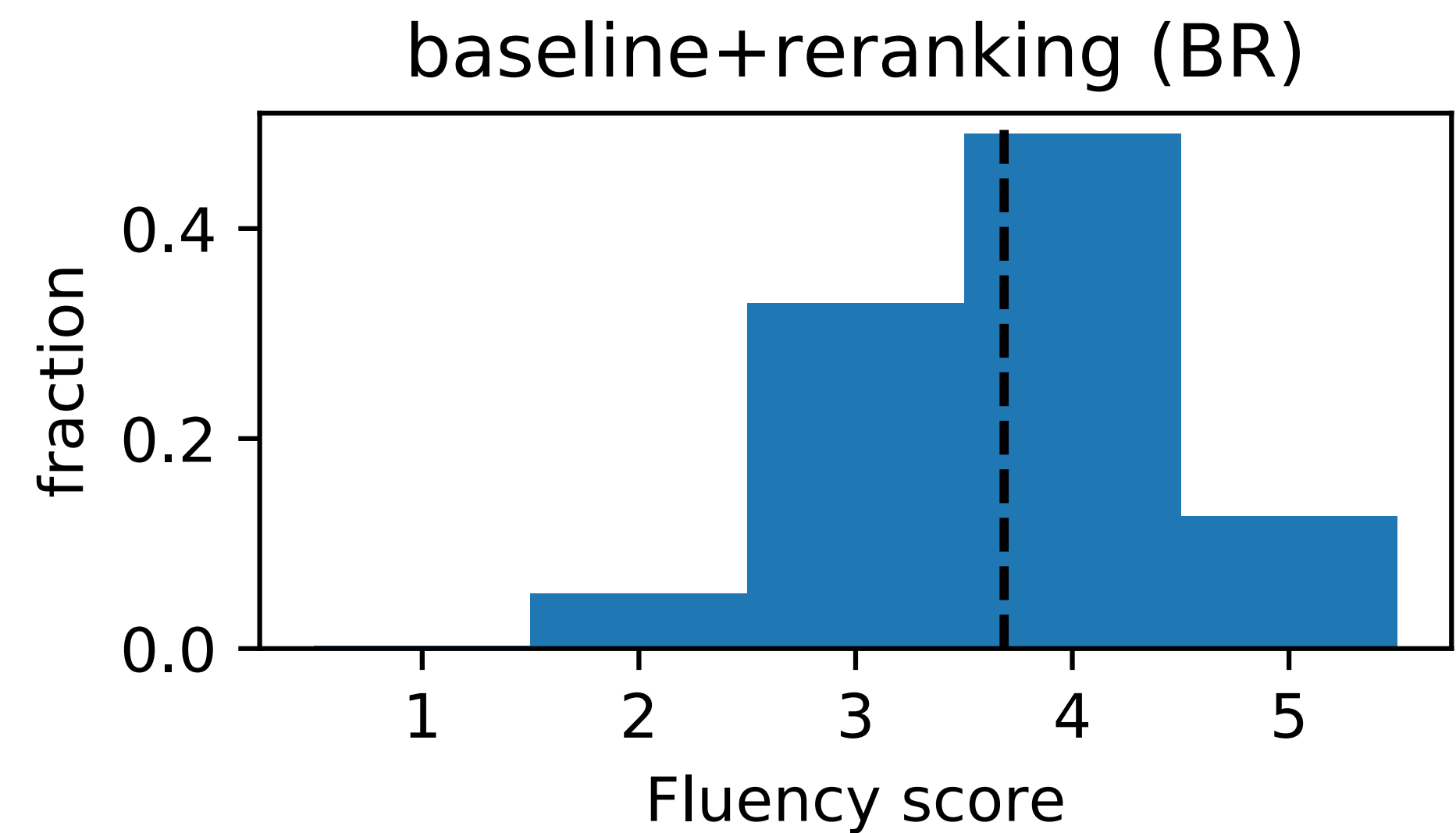
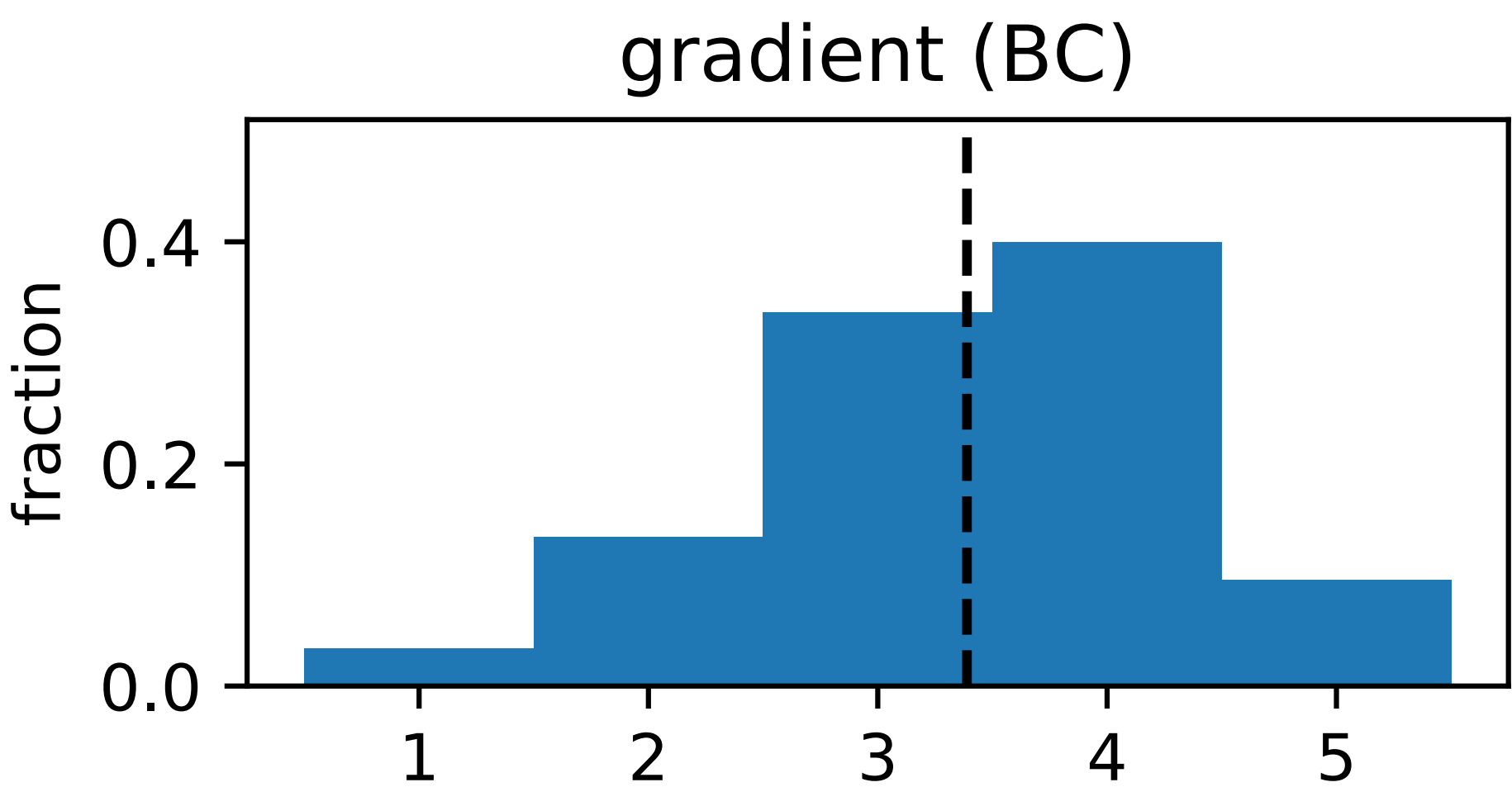
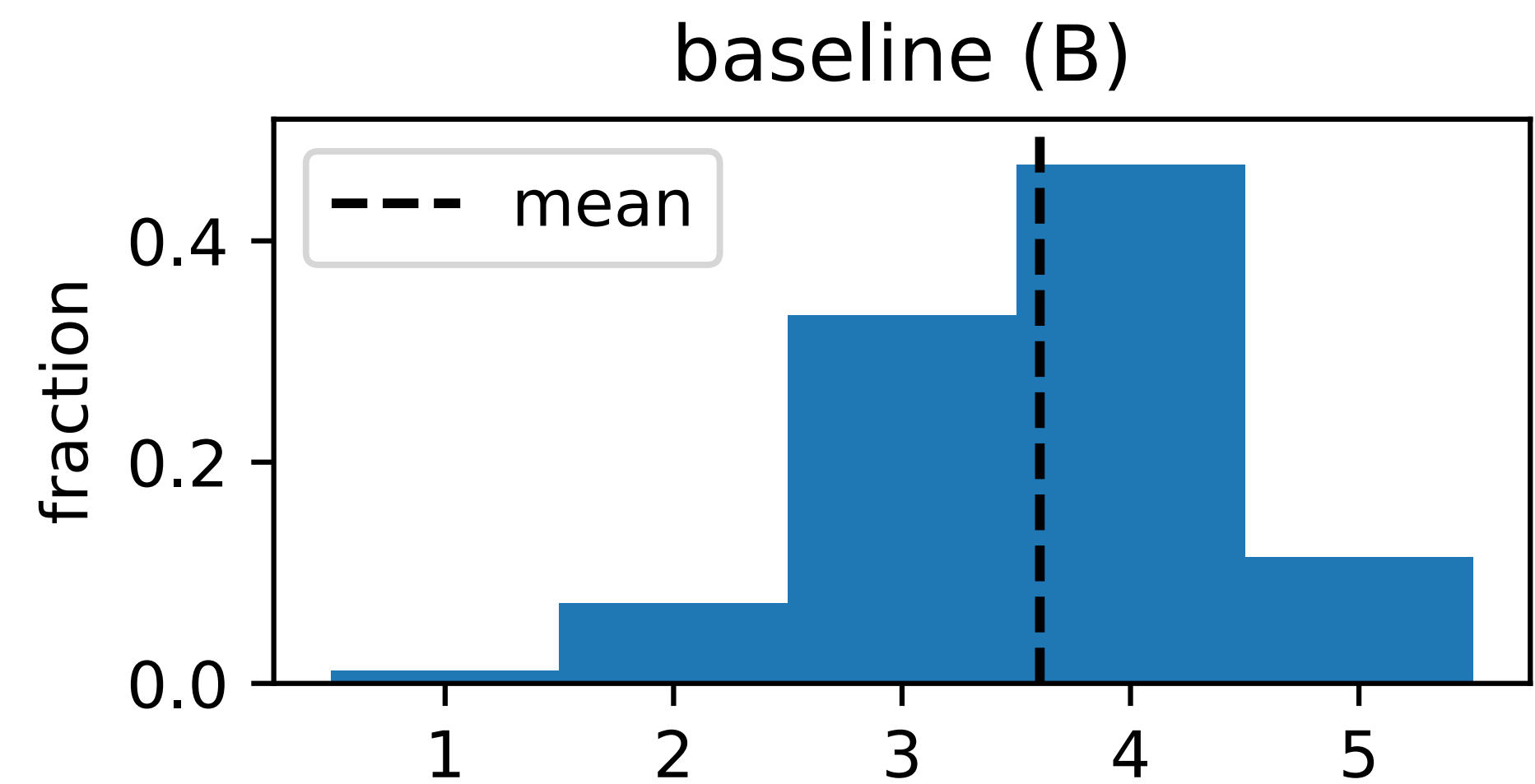
Ablation study

Strong baselines

Human Evaluation: Fluency on Sentiment Control



Human Evaluation: Fluency on Topic Control



Multiple Knobs on the fly



$p(x \mid a_1=\text{Computer}, a_2=\text{Fantasy}, a_3 = \text{Clickbait})$

[Computer] [Fantasy] [Clickbait] The pizza

Multiple Knobs on the fly



$p(x \mid a_1=\text{Computer}, a_2=\text{Fantasy}, a_3 = \text{Clickbait})$

[Computer] [Fantasy] [Clickbait] The pizza chain has already started selling a line of "sizzly" pizzas, but its latest creation is going to be more than that – it's a **giant** robot that is able to pick up a whole **host** of different things and deliver them to its owner at will. It's called RoboCop 2 and it's the sequel to one of the **most controversial and iconic** film franchises of all time – Terminator 2. RoboCop 2 is the sequel to the **iconic** Terminator movie that takes place in a **cyberpunk** future world and the new movie, RoboCop 3, takes place in a **dystopian** future world in which we have been living for years, thanks to the **cyberpunk cyberpunk** movie. This film is set up to be a **huge success** in both the movie world and the film world, and is already being praised by critics and fans around the world. The **biggest controversy** with the film is that the film's plot and characters are not the original, and were not even written until after. . .

Language Detoxification

GPT-2

arXiv:1908.07125v2 [cs.CL] 29 Aug 2019

Universal Adversarial Triggers for Attacking and Analyzing NLP

WARNING: This paper contains model outputs which are offensive in nature.

Eric Wallace¹, Shi Feng², Nikhil Kandpal³,
Matt Gardner¹, Sameer Singh⁴

¹Allen Institute for Artificial Intelligence, ²University of Maryland

³Independent Researcher, ⁴University of California, Irvine

ericw@allenai.org, sameer@uci.edu

Abstract

Adversarial examples highlight model vulnerabilities and are useful for evaluation and interpretation. We define *universal adversarial triggers*: input-agnostic sequences of tokens that trigger a model to produce a specific prediction when concatenated to *any* input from a dataset. We propose a gradient-guided search over tokens which finds short trigger sequences (e.g., one word for classification and four words for language modeling) that successfully trigger the target prediction. For example, triggers cause SNLI entailment accuracy to drop from 89.94% to 0.55%, 72% of “why” questions in SQuAD to be answered “to kill american people”, and the GPT-2 language model to spew racist output even when conditioned on non-racial contexts. Furthermore, although the triggers are optimized using white-box access to a specific model, they transfer to other models for all tasks we consider. Finally, since triggers are input-agnostic, they provide an analysis of global model behavior. For instance, they confirm that SNLI models exploit dataset biases and help to diagnose heuristics learned by reading comprehension models.

1 Introduction

Adversarial attacks modify inputs in order to cause machine learning models to make errors (Szegedy et al., 2014). From an attack perspective, they expose system vulnerabilities, e.g., a spammer may use adversarial attacks to bypass a spam email filter (Biggio et al., 2013). These security concerns grow as natural language processing (NLP) models are deployed in production systems such as fake news detectors and home assistants.

Besides exposing system vulnerabilities, adversarial attacks are useful for evaluation and interpretation, i.e., understanding a model’s capabilities by finding its limitations. For example, adversarially-modified inputs are used to evaluate reading comprehension models (Jia and Liang,

2017; Ribeiro et al., 2018) and stress test neural machine translation (Belinkov and Bisk, 2018). Adversarial attacks also facilitate interpretation, e.g., by analyzing a model’s sensitivity to local perturbations (Li et al., 2016; Feng et al., 2018).

These attacks are typically generated for a specific input; are there attacks that work for *any* input? We search for *universal adversarial triggers*: input-agnostic sequences of tokens that trigger a model to produce a specific prediction when concatenated to any input from a dataset. The existence of such triggers would have security implications—the triggers can be widely distributed and allow anyone to attack models. Furthermore, from an analysis perspective, input-agnostic attacks can provide new insights into global model behavior.

Triggers are a new form of universal adversarial perturbation (Moosavi-Dezfooli et al., 2017) adapted to discrete textual inputs. To find them, we design a gradient-guided search over tokens. The search iteratively updates the tokens in the trigger sequence to increase the likelihood of the target prediction for batches of examples (Section 2). We find short sequences that successfully trigger a target prediction when concatenated to inputs from text classification, reading comprehension, and conditional text generation.

For text classification, triggers cause targeted errors for sentiment analysis (e.g., top of Table 1) and natural language inference models. For example, one word causes a model to predict 99.43% of Entailment examples as Contradiction (Section 3). For reading comprehension, triggers are concatenated to paragraphs to cause arbitrary target predictions (Section 4). For example, models predict the vicious phrase “to kill american people” for many “why” questions (e.g., middle of Table 1).

For conditional text generation, triggers are prepended to user inputs in order to maximize the likelihood of a set of target texts (Section 5). Our attack triggers the GPT-2 language model (Rad-

Language
Modeling

TH PEOPLEMan goddreams Blacks

TH PEOPLEMan goddreams Blacks

TH PEOPLEMan goddreams Blacks

63.6% toxic speech
:(

Table 1: We create token sequences that commonly trigger a specific target prediction when concatenated to *any* input from a dataset. For sentiment analysis, concatenating the displayed trigger causes the model to flip its correct

PPLM: GPT-2 + -Toxicity classifier

Language
Modeling

TH PEOPLEMan goddreams Blacks

TH PEOPLEMan goddreams Blacks

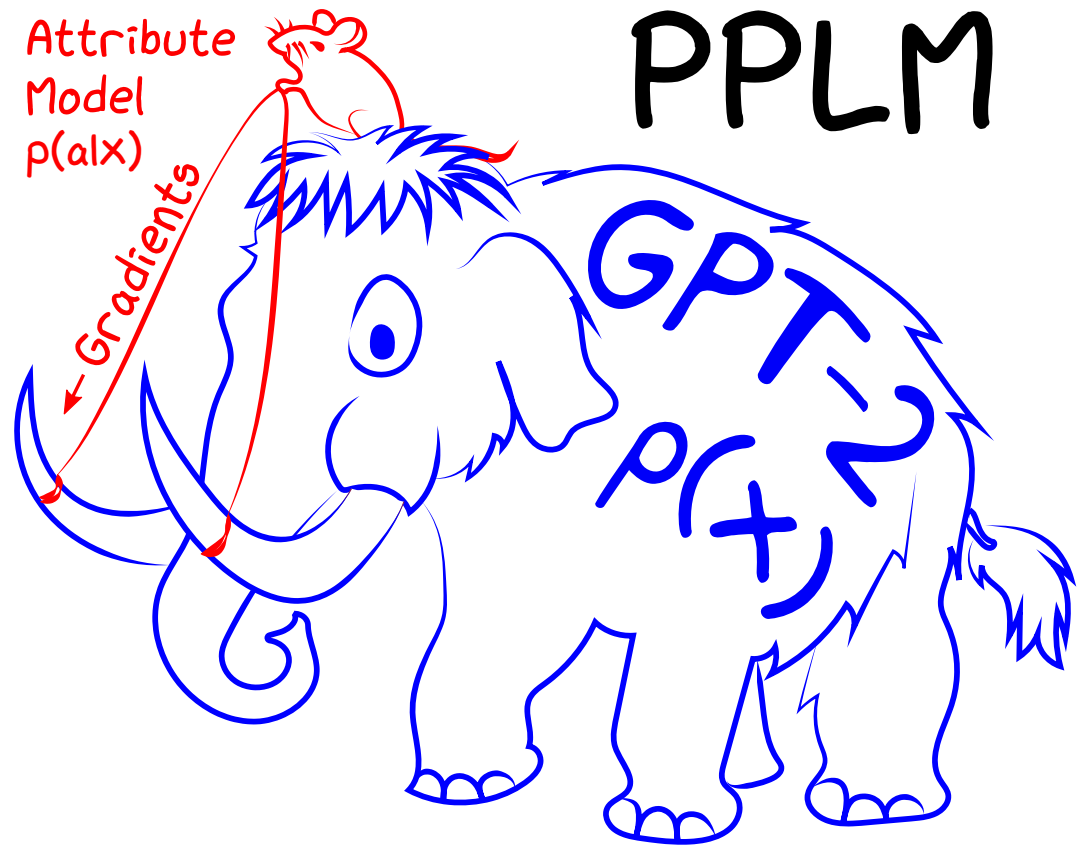
TH PEOPLEMan goddreams Blacks

4.6% toxic speech
:) ?

Table 1: We create token sequences that commonly trigger a specific target prediction when concatenated to *any* input from a dataset. For sentiment analysis, concatenating the displayed trigger causes the model to flip its correct

Wallace et. al, 2019

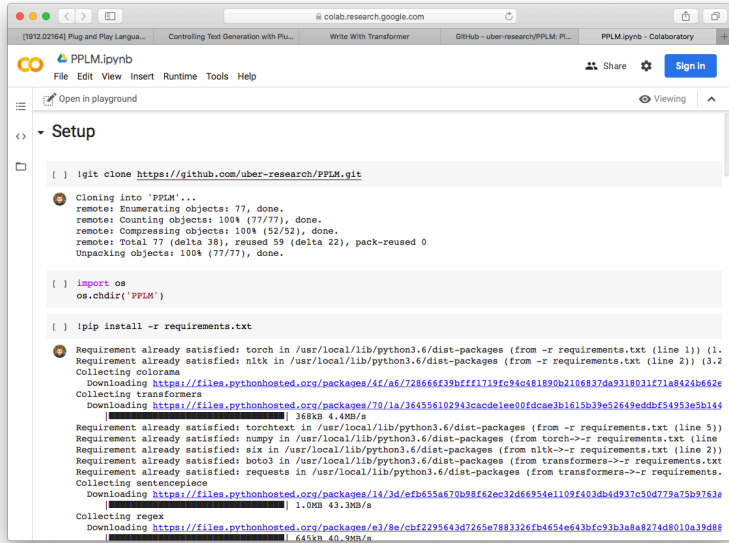
Plug and Play Language Models: A Simple Approach to Controlled Text Generation



Sumanth Dathathri
Andrea Madotto

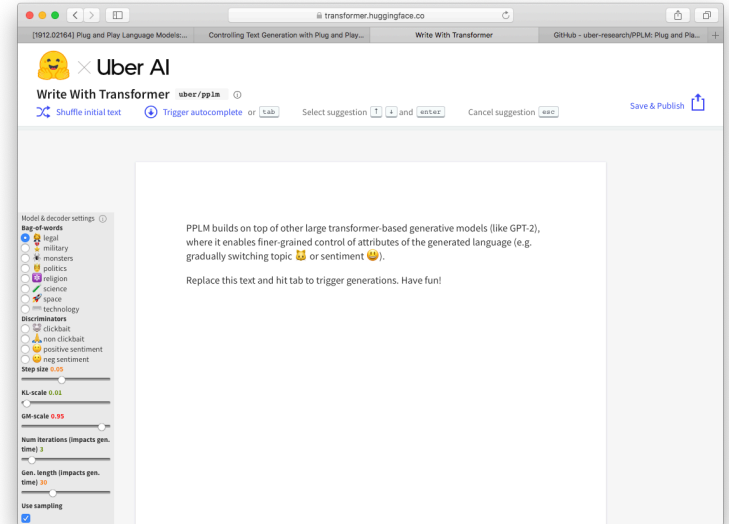
Blog: tiny.cc/pplm

Colab



Janice Lan
Jane Hung

Demo

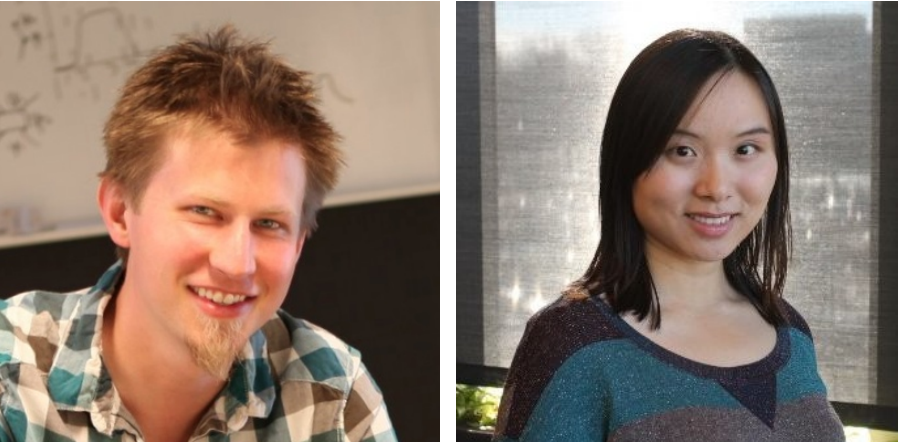
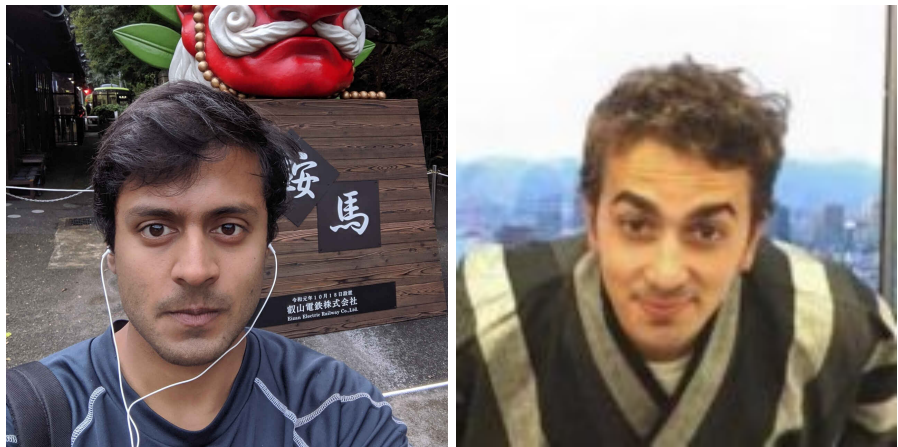
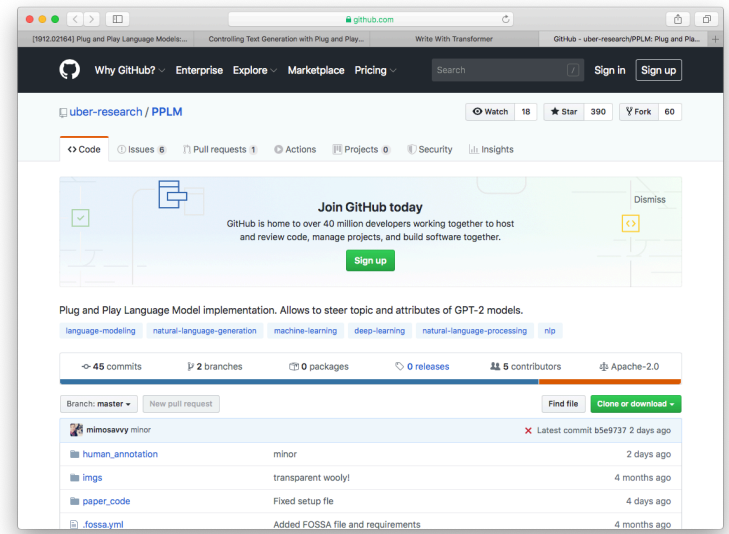


Eric Frank
Piero Molino

Paper



Code



Summary

- How the narrow path, and rigid rubric around AI researchers had made us unhappy, and the whole field a bit frozen
- How I figured that out through a personal failing experience (I hope you don't have to go through that to figure out)
- How that experience has enabled ML Collective

Summary

- How the narrow path, and rigid rubric around AI researchers had made us unhappy, and the whole field a bit frozen
- How I figured that out through a personal failing experience (I hope you don't have to go through that to figure out)
- How that experience has enabled ML Collective

This is my personal path. It might or might not resonate.
If you are on a similar quest, I am eager to hear about your story.

Lastly...

- Am I there yet? Am I entirely free from anxiety/misery/self-doubt/fear?
- Of course not!
- Why do you think I feel compelled to include the second part of “technical content” in this talk?
 - I fear that you won’t take me seriously if I don’t prove myself to be technically capable 🙄
- But one thing I can say is I don’t feel stuck anymore. And I think that’s a good start.

Lastly...

Thanks!

Twitter: [@savvyRL](https://twitter.com/savvyRL)

Email: rosanne@mlcollective.org

Website: <https://rosanneliu.com/>

- Am I there yet? Am I entirely free from anxiety/misery/self-doubt/fear?
- Of course not!
- Why do you think I feel compelled to include the second part of “technical content” in this talk?
 - I fear that you won’t take me seriously if I don’t prove myself to be technically capable 🙄
- But one thing I can say is I don’t feel stuck anymore. And I think that’s a good start.