

# Diverse and Efficient Ensembling of Deep Networks

Alexandre Ramé

July 2<sup>nd</sup> 2024 at CAP-RFIAP in Lille

## Jury of the PhD defense (October 11<sup>th</sup> 2023):

Pr. Graham Taylor, *University of Guelph & Vector Institute*

Pr. Christian Wolf, *Naver Labs*

DR. Cordelia Schmid, *INRIA & Google*

DR. Léon Bottou, *Meta AI*

Dr. Thomas Wolf, *HuggingFace*

Pr. Patrick Gallinari, *Sorbonne Université & Criteo*

**Thesis director:** Pr. Matthieu Cord, *Sorbonne Université & valeo.ai*

# Artificial intelligence revolution

AlphaGo

Board games



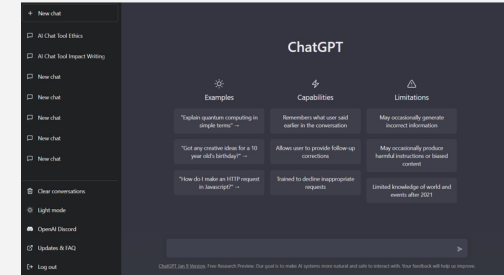
Stable  
diffusion

Image generation

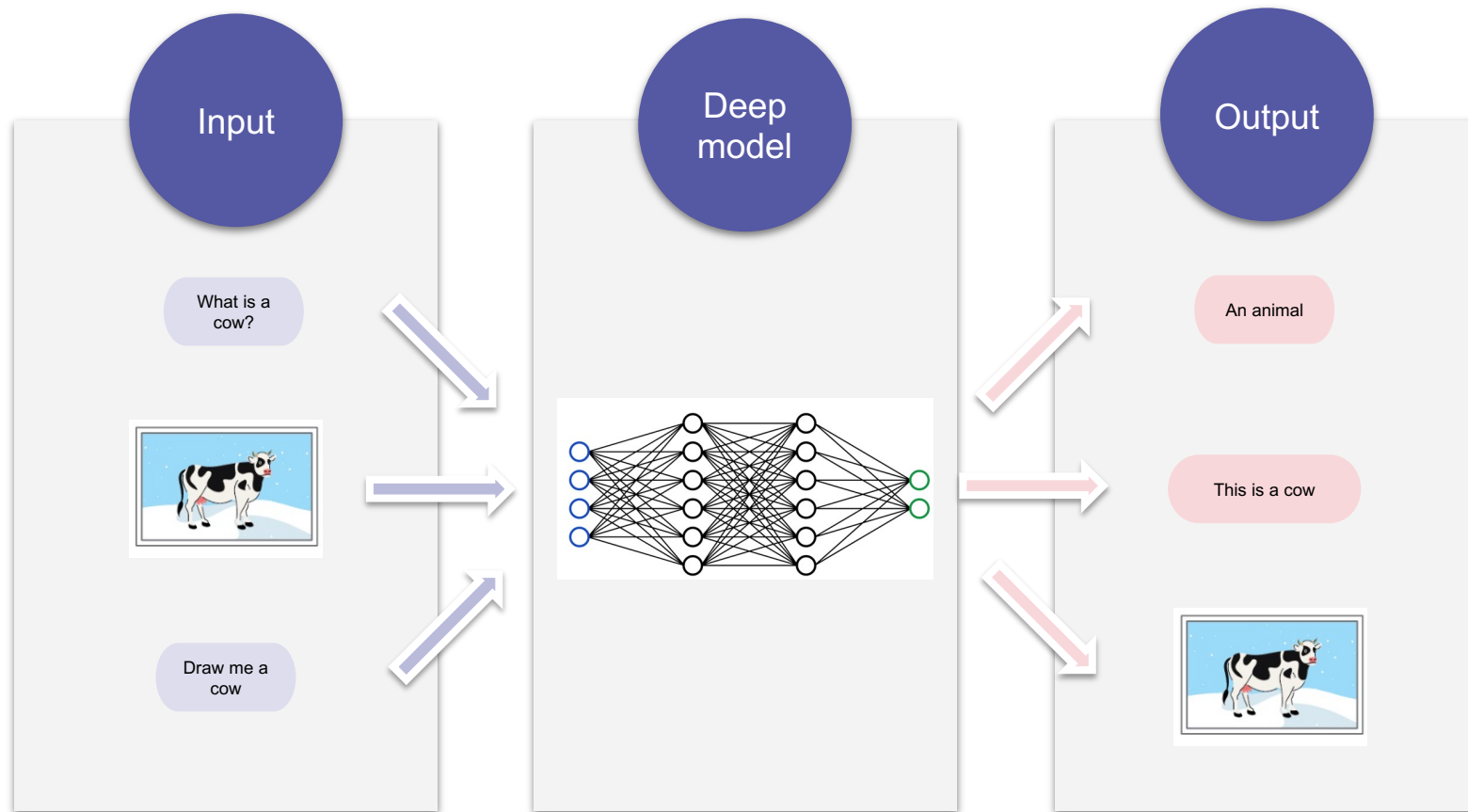


ChatGPT

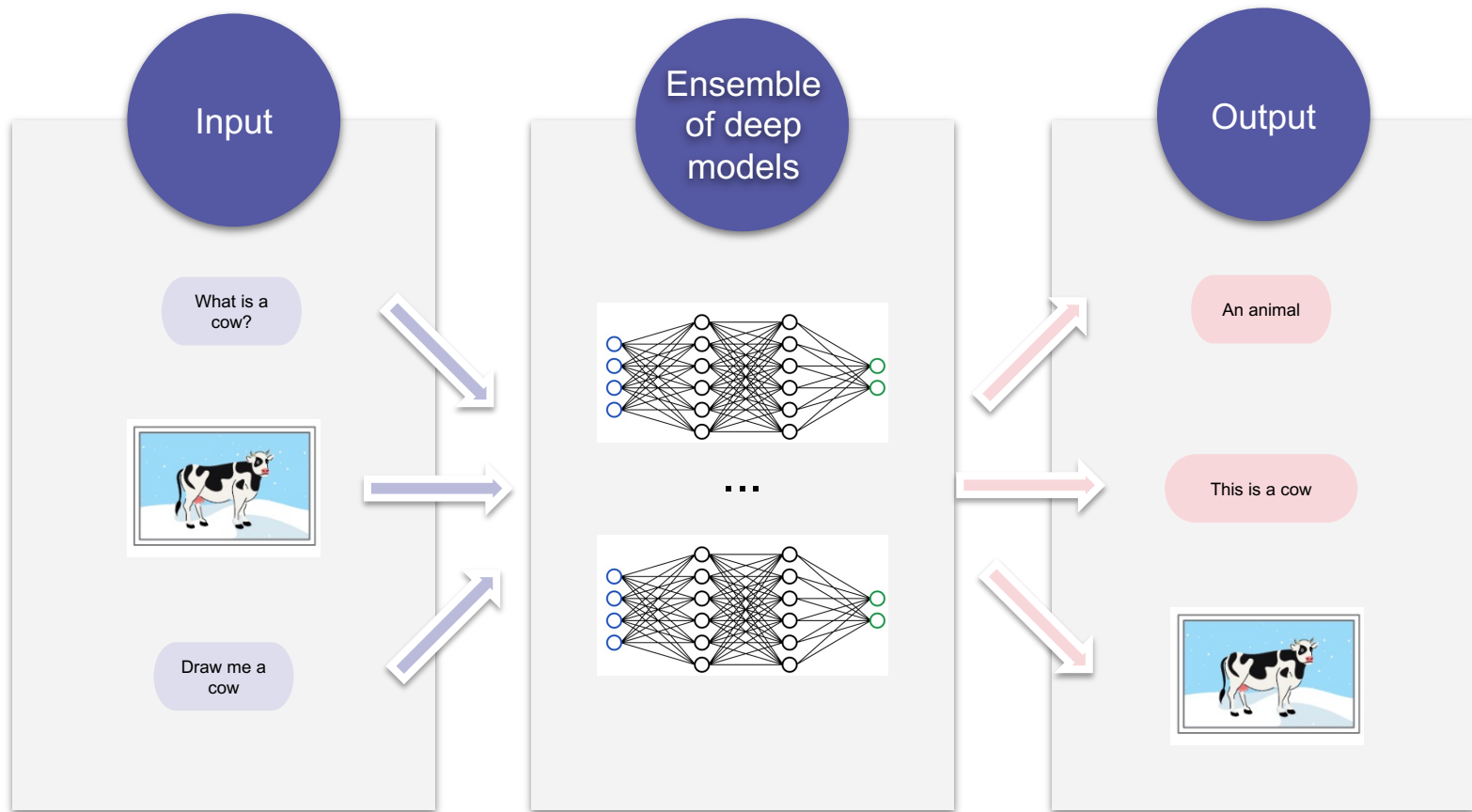
Conversational agent



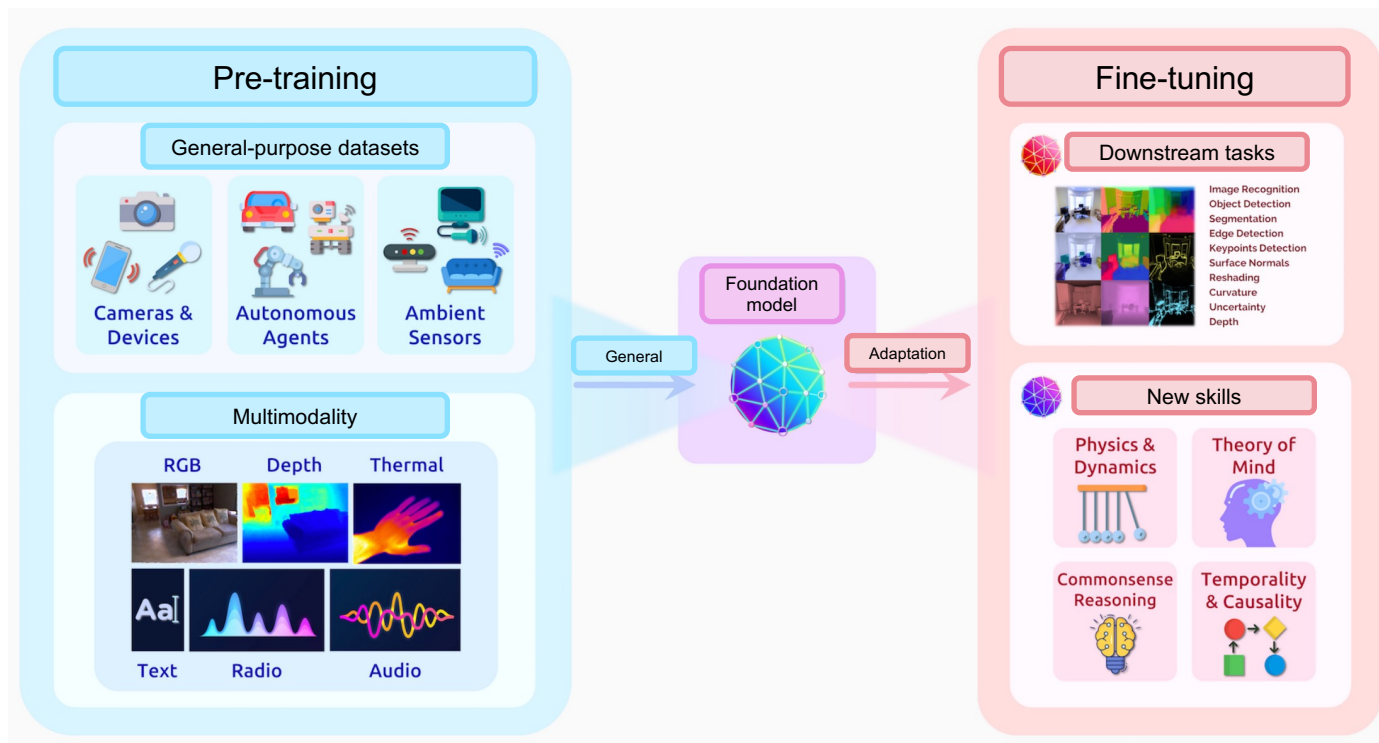
# Deep learning



# Ensembling in deep learning



# Transfer learning and fine-tuning from foundation model



# Plan

---

## Part I.

Weight averaging for out-of-distribution generalization.

## Part II.

Weight averaging for reinforcement learning from human feedback

# Plan

---

## Part I.

Weight averaging for out-of-distribution generalization.

## Part II.

Weight averaging for reinforcement learning from human feedback



DiWA: diverse weight averaging for out-of-distribution generalization.

**Alexandre Ramé**, Matthieu Kirchmeyer, Thibaud Rahier, Alain Rakotomamonjy, Patrick Gallinari, Matthieu Cord. NeurIPS 2022.



Model ratatouille: recycling diverse models for out-of-distribution generalization.

**Alexandre Ramé**, Kartik Ahuja, Jianyu Zhang, Matthieu Cord, Léon Bottou and David Lopez-Paz. ICML 2023.



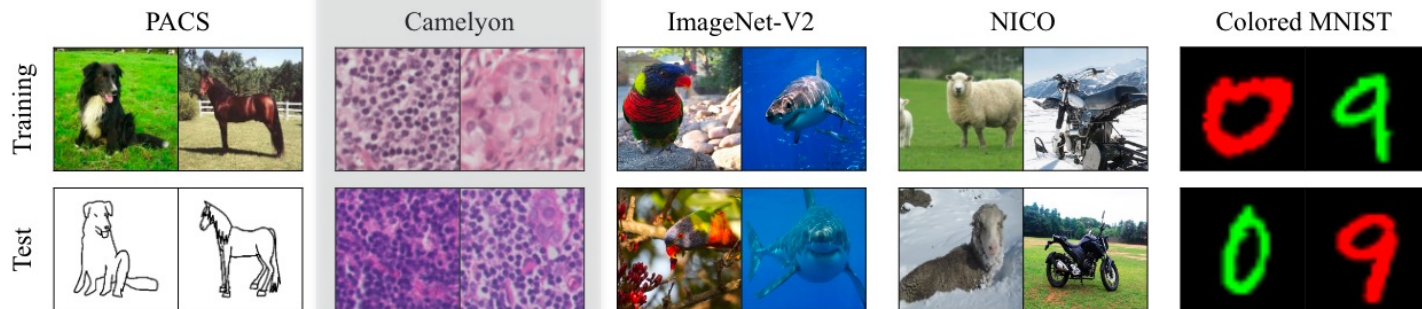
# Generalization on test samples

## Challenge

Generalization, notably under distribution/domain shift.

## Importance

Robustness for responsible and fair use.



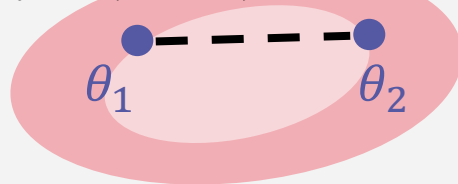
(cancer detection, with different hospitals in train and test)



# Model merging

We investigate strategies to merge two models  $\theta_1$  and  $\theta_2$  in the weight space (despite the non-linearities).

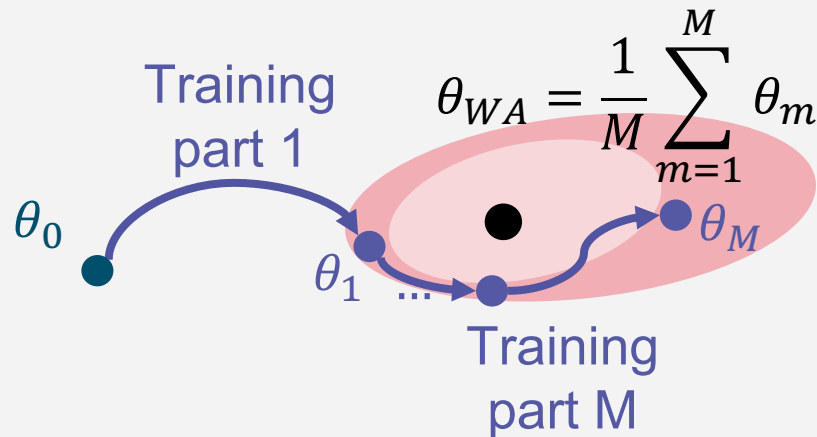
$$\theta_\lambda = (1 - \lambda) \cdot \theta_1 + \lambda \cdot \theta_2$$



Weight averaging = simple & efficient ensembling method deep models.

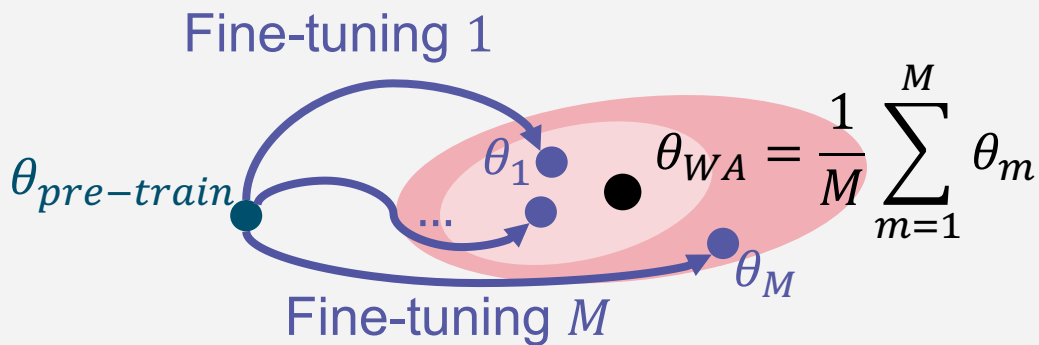
# Weight averaging along a training trajectory

Moving average [Izmailov2018]:  
checkpoints collected along a training  
trajectory remain linearly connected.



# Weight averaging from multiple trajectories

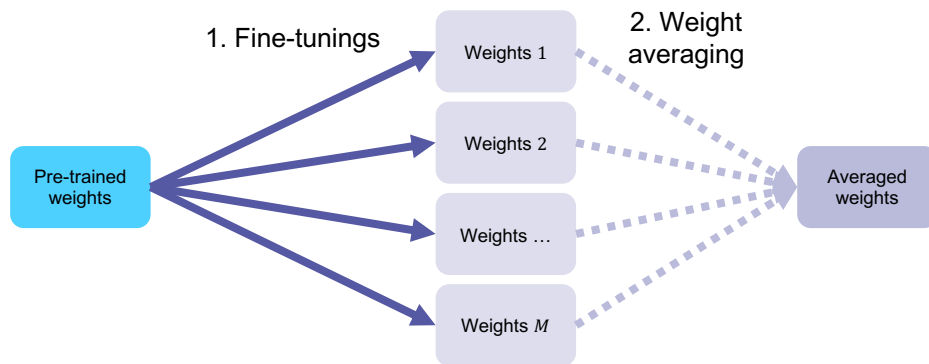
When fine-tuned from a shared pre-trained model, weights remain linearly connected.



# DiWA recipe

From a shared pre-trained network:

1. Launch multiple runs with different hyperparameters (like a grid search).
2. Weight average all fine-tuned models (rather than selecting the best one).

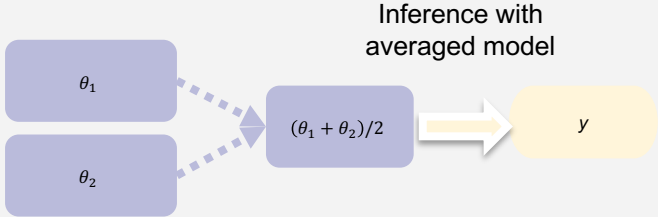
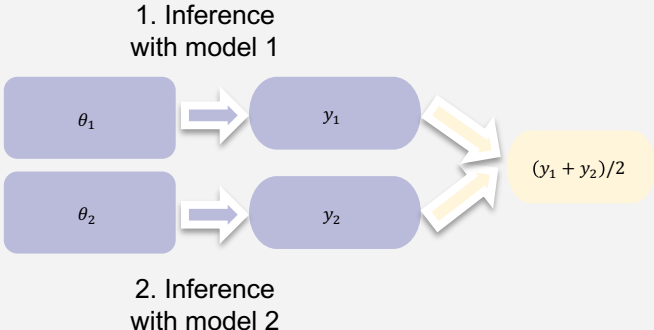


DiWA: diverse weight averaging for out-of-distribution generalization.

**Alexandre Ramé**, Matthieu Kirchmeyer, Thibaud Rahier, Alain Rakotomamonjy, Patrick Gallinari, Matthieu Cord. NeurIPS 2022.

[Wortsman2022] Model soups: averaging weights of multiple fine-tuned models improves accuracy. ICML.

# Weight averaging as efficient & improved ensembling

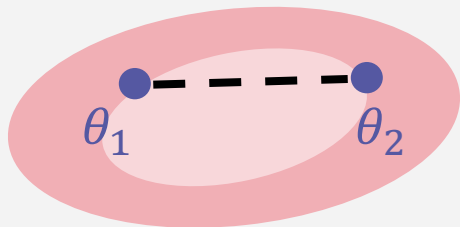
Name	Weight averaging	Prediction averaging (traditional ensembling)
What	 <p>Inference with averaged model</p>	 <p>1. Inference with model 1</p> <p>2. Inference with model 2</p>
Inference cost	1 single forward	2 forwards
Constraint	Weights linearly mode connected for a given architecture	No constraint
Distribution shifts	Generalization by variance reduction	Generalization by variance reduction
Label corruption	Reduced memorization by removing run-specific features	Memorization of corrupted labels

# The 3 criteria for successful weight averaging

---

## Linear connectivity

The weights should remain linearly connected.



## Individual accuracies

The weights should be individually accurate.

$\{\theta_i\}_i$  accurate  
independently

## Diversity

The predictions should be sufficiently diverse.

$$\theta_1 \neq \theta_2$$

# The 3 sources of diversity in DiWA

---

## Hyper-parameters

- Learning rate.
- Weight decay.
  - etc.

## Data

- Batch orders.
- Bagging.

## Random factors

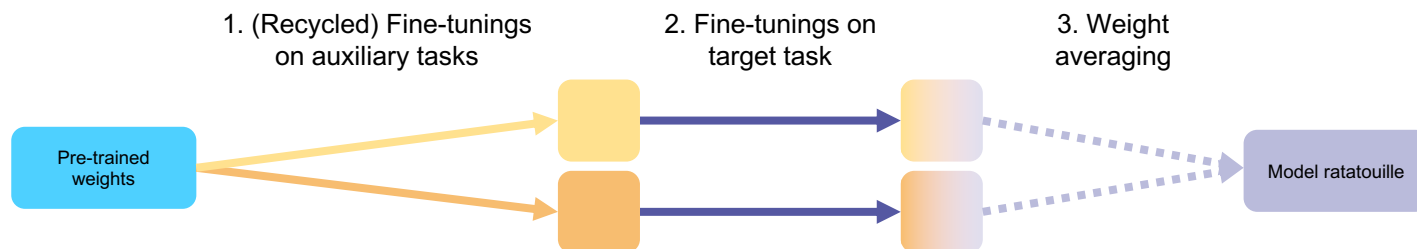
- Dropout.
- Augmentation.
- Learning stochasticity.

# Ratatouille recipe



From a shared pre-trained network:

1. Recycle multiple fine-tunings on auxiliary tasks.
2. Launch multiple fine-tunings on the target task with different initializations.
3. Average all the fine-tuned weights.



Model ratatouille: recycling diverse models for out-of-distribution generalization.

**Alexandre Ramé**, Kartik Ahuja, Jianyu Zhang, Matthieu Cord, Léon Bottou and David Lopez-Paz. ICML 2023.



# Does Ratatouille meet the 3 criteria for weight averaging ? —

## Linear connectivity

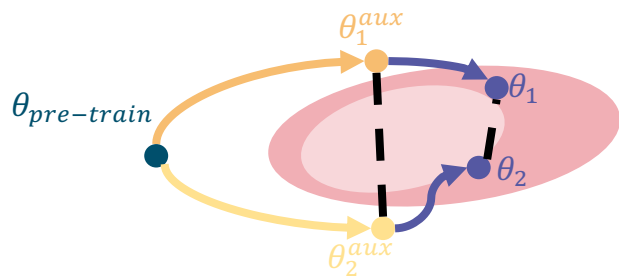
Yes, weights remain linearly connected even when starting from diverse inits.

## Individual accuracies

Yes, by transferring rich features on auxiliary tasks.

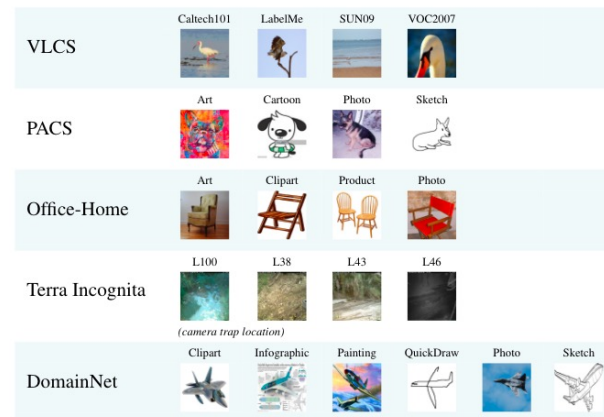
## Diversity

Yes !! huge gain in diversity caused by different inits.



# New state of the art on DomainBed

Algo	VLCS	PACS	OfficeH	TerraInc	DNet	Average
ERM	77.5	85.5	66.5	46.1	40.9	63.3
MA	78.2	87.5	70.6	50.3	46.9	66.5
<u>DiWA</u>	<u>78.4</u>	<u>88.7</u>	<u>72.1</u>	<u>51.4</u>	<u>47.4</u>	<u>67.6</u>
<b>Ratatouille</b>	<b>78.5</b>	<b>89.5</b>	<b>73.1</b>	<b>51.8</b>	<b>47.5</b>	<b>68.1</b>



# Plan

---

## Part I.

Weight averaging for out-of-distribution generalization.

## Part II.

Weight averaging for reinforcement learning from human feedback



Rewarded soups: towards Pareto-optimal alignment by interpolating weights fine-tuned on diverse rewards.

**Alexandre Ramé**, Guillaume Couairon, Corentin Dancette, Jean-Baptiste Gaya, Mustafa Shukor, Laure Soulier, Matthieu Cord. NeurIPS 2023.



WARM: On the Benefits of Weight Averaged Reward Models.

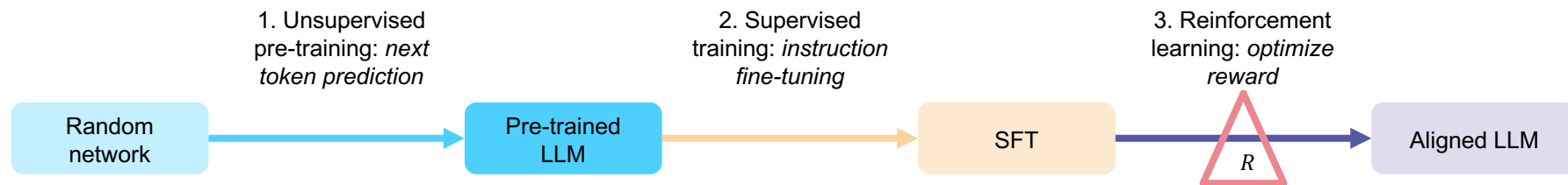
**Alexandre Ramé**, Nino Vieillard, Léonard Hussenot, Robert Dadashi, Geoffrey Cideron, Olivier Bachem, Johan Ferret. ICML 2024.



WARP: On the Benefits of Weight Averaged Rewarded Policies.

**Alexandre Ramé**, Johan Ferret, Nino Vieillard, Robert Dadashi, Léonard Hussenot, Pierre-Louis Cedoz, *et al.*. arXiv 2024.

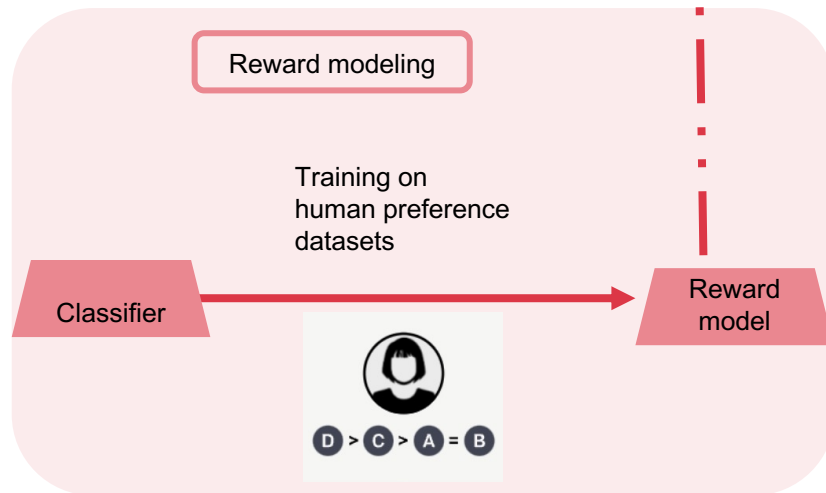
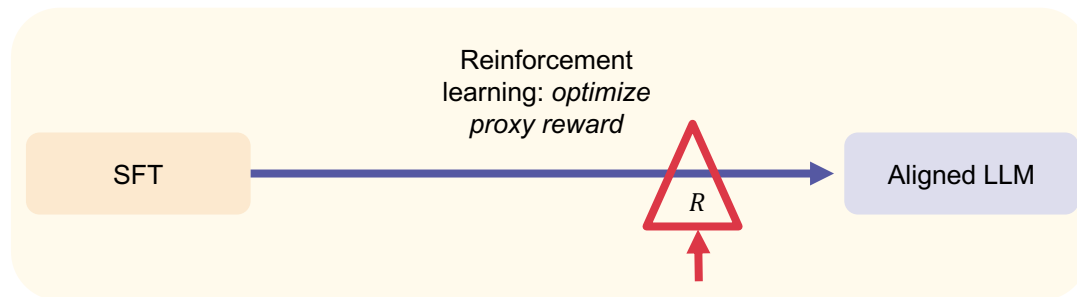
# Fine-tunings to align LLMs with human expectations



## Why RL:

- Evaluates the **sentence** rather than tokens independently.
- Does not require supervised samples, but instead a **reward**.
- More **online exploration**.

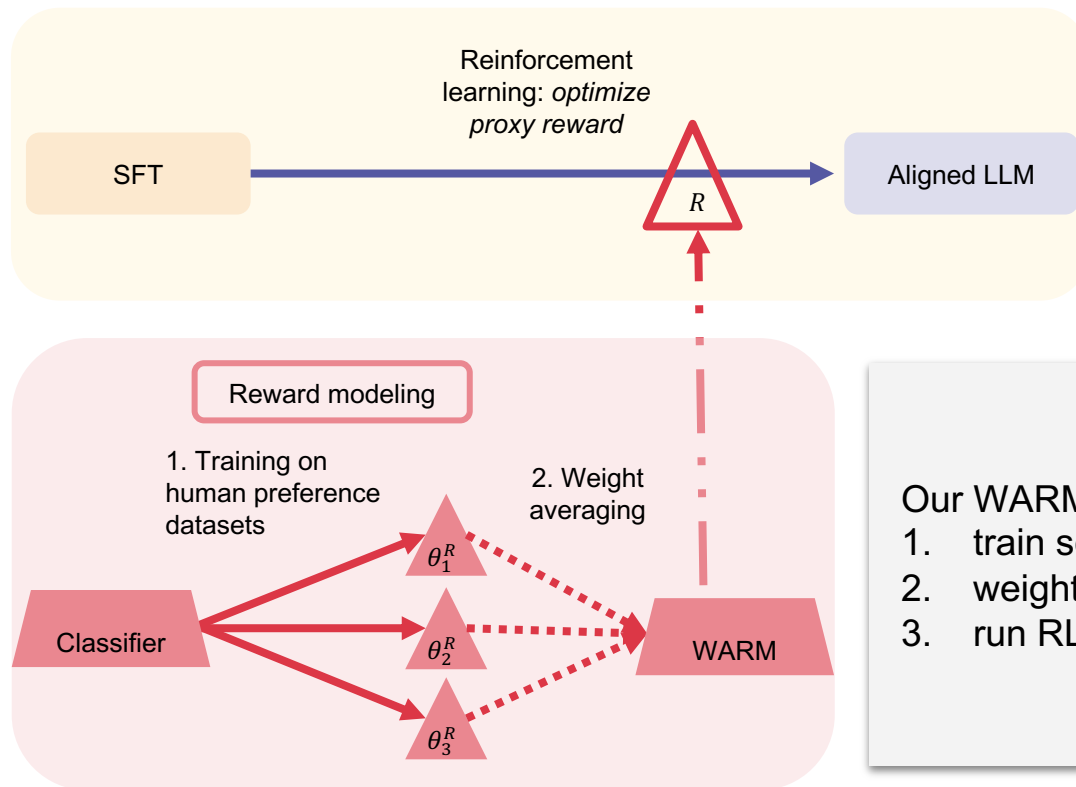
# Reward modeling challenge



Problem: true reward not available.

Challenge: designing **reliable and robust** proxy reward models.

# WARM: Weight Averaged Reward Models



Our WARM solution:

1. train several reward models,
2. weight average them,
3. run RLHF to optimize it.



WARM: On the Benefits of Weight Averaged Reward Models.

**Alexandre Ramé**, Nino Vieillard, Léonard Hussenot, Robert Dadashi, Geoffrey Cideron, Olivier Bachem, Johan Ferret. ICML 2024.



# Diversity of opinions

---

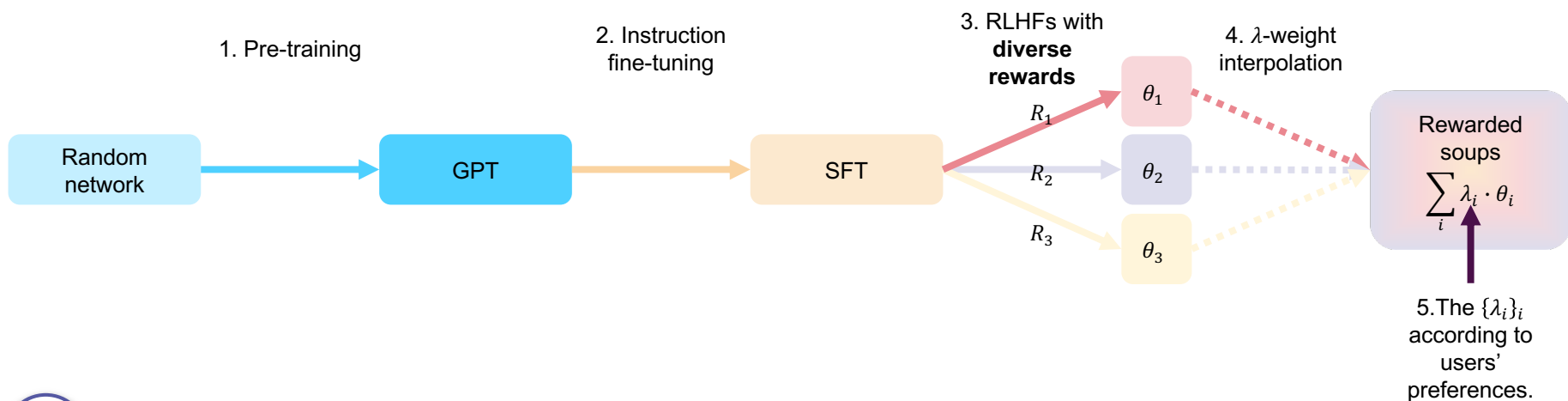
Humans have **diverse opinions** (politics, aesthetics, etc) and **different expectations** from machines (helpfulness vs. harmlessness),  
leading to **fairness** and **engineering** issues:  
“human aligned artificial intelligence is a multi-objective problem”.



# Rewarded soups recipe



1. From a shared pre-trained foundation model,
2. Fine-tuned to follow instructions,
3. Launch one RL fine-tuning for each proxy reward, each representing an opinion,
4. Interpolate the weights specialized on diverse rewards,
5. Reveal the front of solutions (and select one interpolating coefficient).



Rewarded soups: towards Pareto-optimal alignment by interpolating weights fine-tuned on diverse rewards.

Alexandre Ramé, Guillaume Couairon, Corentin Dancette, Jean-Baptiste Gaya, Mustafa Shukor, Laure Soulier, Matthieu Cord. NeurIPS 2023.



# Summarization: completeness vs. faithfulness

## Hillary Clinton email controversy

FBI Director James Comey told Congress on Sunday a recent review of newly discovered emails did not change the agency's conclusion reached in July that no charges were warranted in the case of Hillary Clinton's use of a private email server. U.S. Republican Representative Jason Chaffetz said in a tweet that Comey had informed him of the conclusion. Comey's letter to Congress informing it of the newly discovered emails had thrown Clinton's presidential race against Republican Donald Trump into turmoil.

FBI tells Congress it has not changed its conclusion from July that no charges are warranted in the Hillary Clinton email server case, but has recently discovered new emails related to the investigation.

$R_1$ : completeness

$\theta_1$

Pre-trained  
LLM

SFT

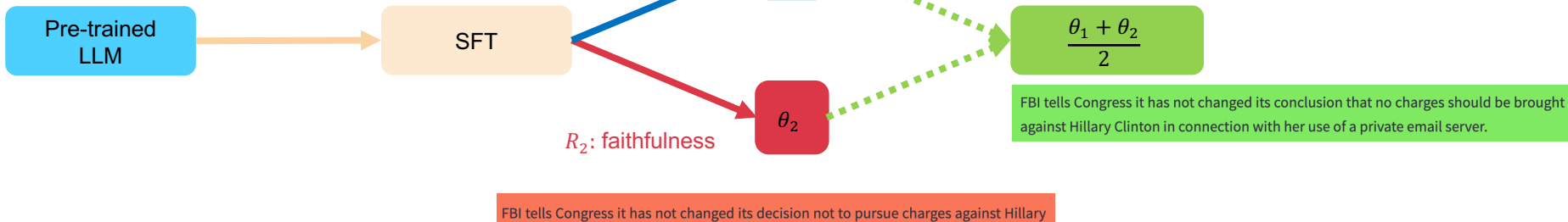
$$\frac{\theta_1 + \theta_2}{2}$$

$R_2$ : faithfulness

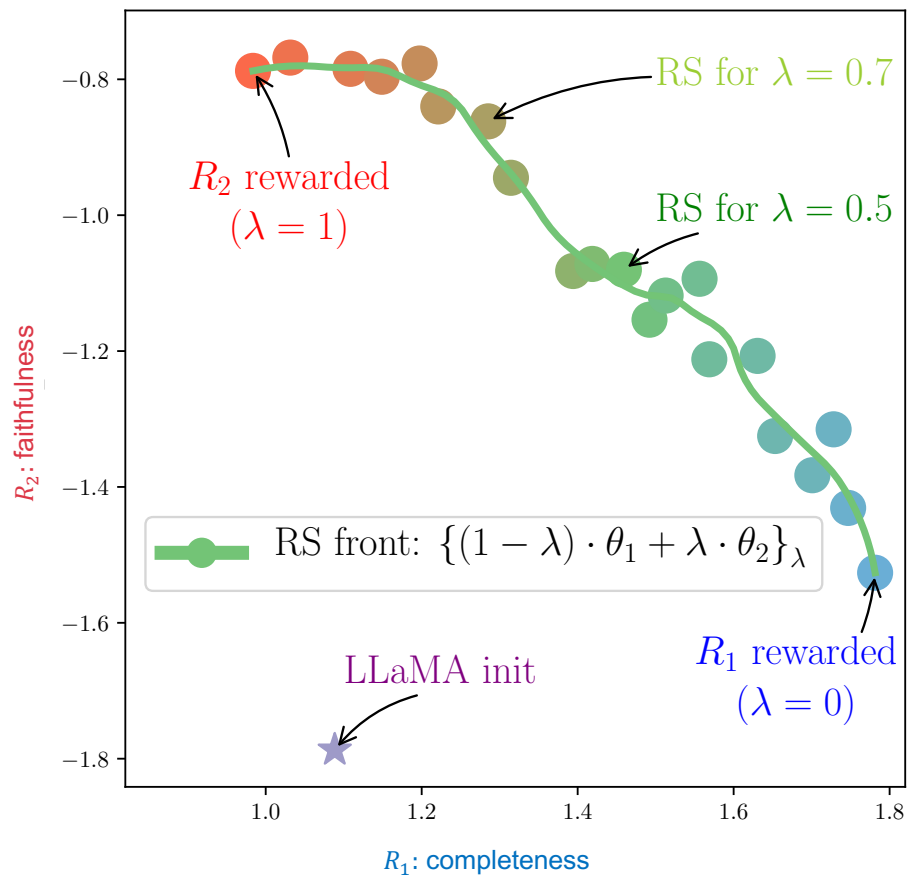
$\theta_2$

FBI tells Congress it has not changed its conclusion that no charges should be brought against Hillary Clinton in connection with her use of a private email server.

FBI tells Congress it has not changed its decision not to pursue charges against Hillary



# Pareto front of solutions



# Rewarded soups in multiple setups

---

## Text

- Summarization (news, reddit).
- Conversational assistant.
  - Technical Q&As.
- Movie review generation.

## Multimodal

- Image captioning.
- Image generation with diffusion models.
- Visual grounding.
- Visual question answering.

## Locomotion

- Robot continuous control.

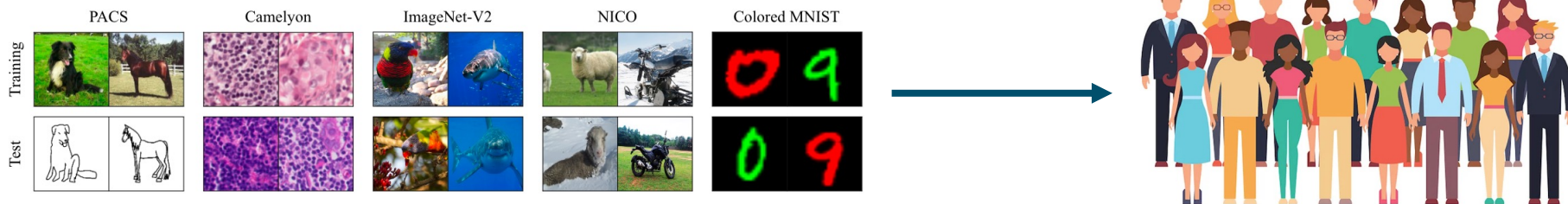
# Conclusion

Summary of contributions and perspectives

# 1<sup>st</sup> contribution: **diversity** for robust ensembling

Combining diverse members as a robust strategy to handle train-test differences.

- Distribution shifts for out-of-distribution generalization.
- Reward misspecification for alignment.

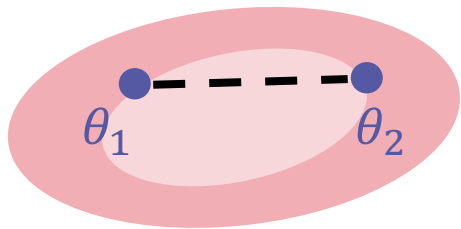


## 2<sup>nd</sup> contribution: weight averaging for **efficient** ensembling —

Linear mode connectivity verified in all considered scenarios:

- Multiple setups: supervised and reinforcement learning.
- Multiple tasks: classification or generation.
- Multiple modalities: text and image.

And thus weight averaging as a practical strategy for training LLMs.



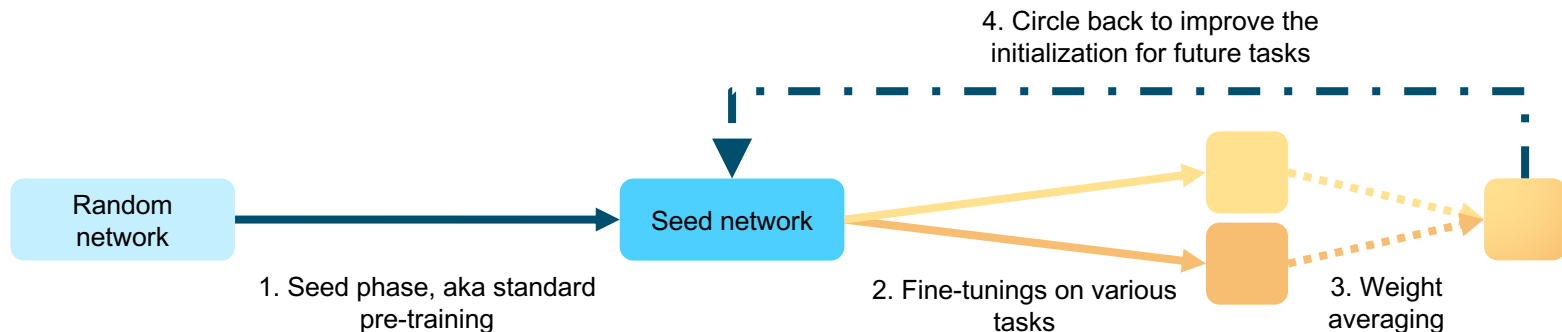
The larger the model,  
the easier the weight averaging.



## 3<sup>th</sup> contribution: large-scale fine-tuning

Scale fine-tuning like pre-training was scaled for improved results and alignment:

1. Pre-training of foundation models.
2. Parallelizable fine-tunings on various tasks.
3. Weight averaging to combine information.
4. Iterate

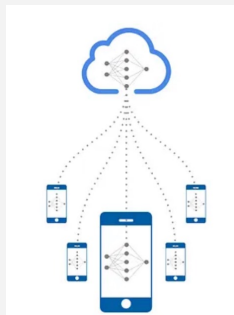


# Towards updatable machine learning

---

## Federated learning

Only share weights, data remain private.



## Distributed learning

Embarrassingly simple parallelization with multiple independent trainings.

## Open source

ML come with risks of centralization, and a two-speed research.

⇒ collaborative solutions.

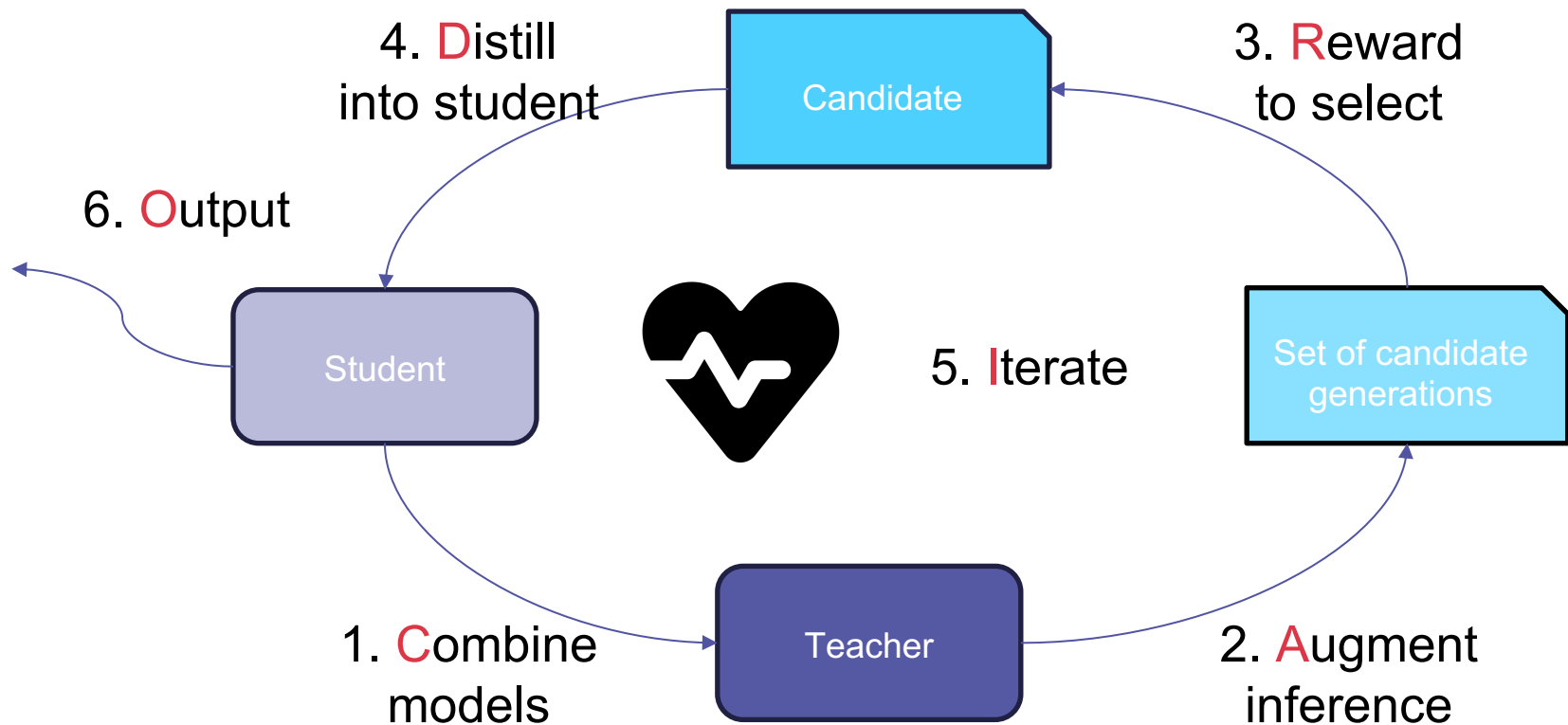


[Raffel2023] A Call to Build Models Like We Build Open-Source Software. ACM.

[Douillard2023] DiLoCo: Distributed Low-Communication Training of Language Models .



# Towards iterated amplification: CARDIO framework



## Final perspective

---

Investigate how model merging can align towards better and safer models.

# Thank you !

Paper	Title	Conference	Year
DICE	Diversity in deep ensembles via conditional redundancy adversarial estimation	ICLR	2021
MixMo	Mixing multiple inputs for multiple outputs via deep subnetworks	ICCV	2021
Fishr	Invariant gradient variances for out-of-distribution generalization	ICML	2022
DiWA	Diverse weight averaging for out-of-distribution generalization	NeurIPS	2022
Ratatouille	Recycling diverse models for out-of-distribution generalization	ICML	2023
Rewarded soups	Towards Pareto-optimal alignment by interpolating weights	NeurIPS	2023
WARM	On the benefits of weight averaged reward models	ICML	2024
WARP	On the benefits of weight averaged rewarded policies	arXiv	2024
MixShare	Towards efficient feature sharing in MIMO architectures	CVPR W	2022
DyTox	Transformers for continual learning with dynamic token expansion	CVPR	2022
Interpolate	Pre-train, fine-tune, interpolate: a three-stage strategy for generalization	NeurIPS W	2022
UniVAL	Unified model for image, video, audio and language tasks	TMLR	2023
EvAlign	Evaluating and reducing the flaws of LMMs with in-context-learning?	ICLR	2024
DAP	Direct Language Model Alignment from Online AI Feedback	arXiv	2024
Gemma2	Improving Open Language Models at a Practical Size	Technical report	2024

