Diverse and Efficient Ensembling of Deep Networks

Thesis Defense – Alexandre Ramé

October 11th 2023

Jury:
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*
Invitee: Dr. David Lopez-Paz, *Meta AI*
Thesis director: Pr. Matthieu Cord, *Sorbonne Université & valeo.ai*

SORBONNE UNIVERSITÉ

ISIR

MLiA Machine Learning & Deep Learning for Information Access

Artifical intelligence revolution

AlphaGo

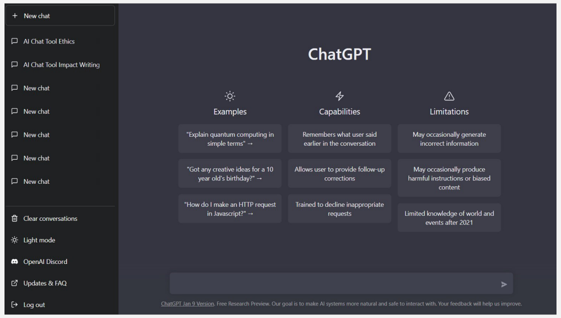
Board games

Stable diffusion

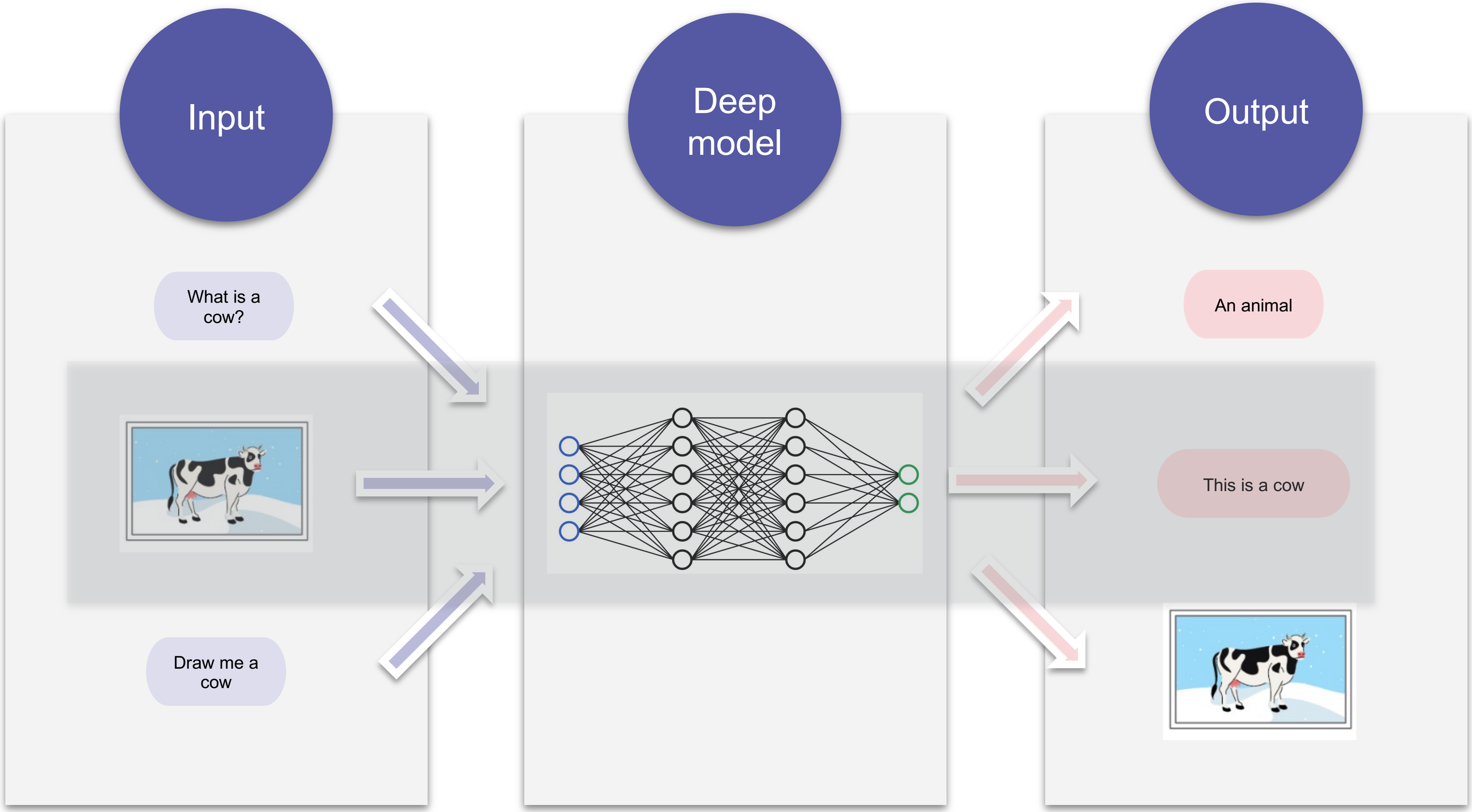
Image generation

ChatGPT

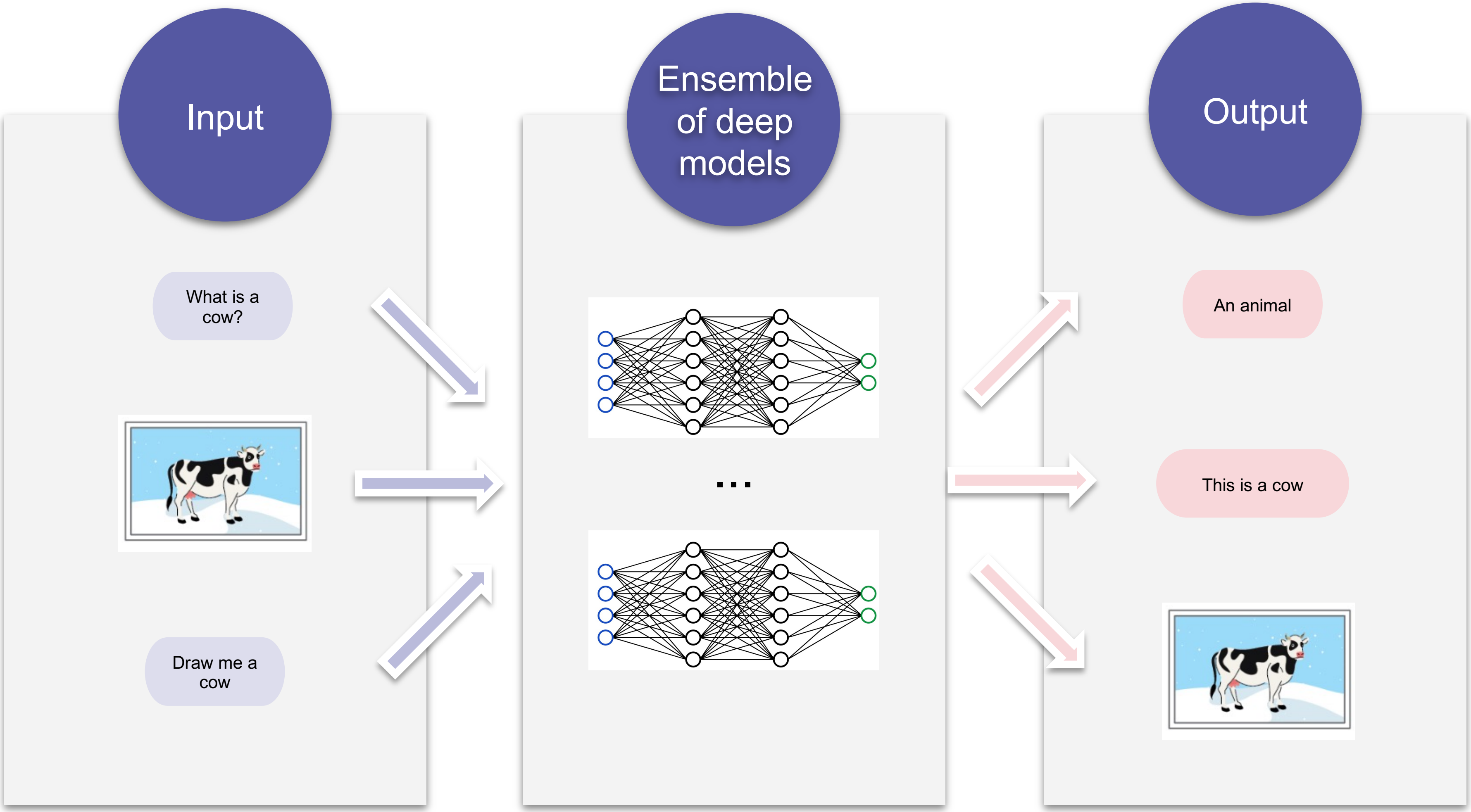
Conversational agent



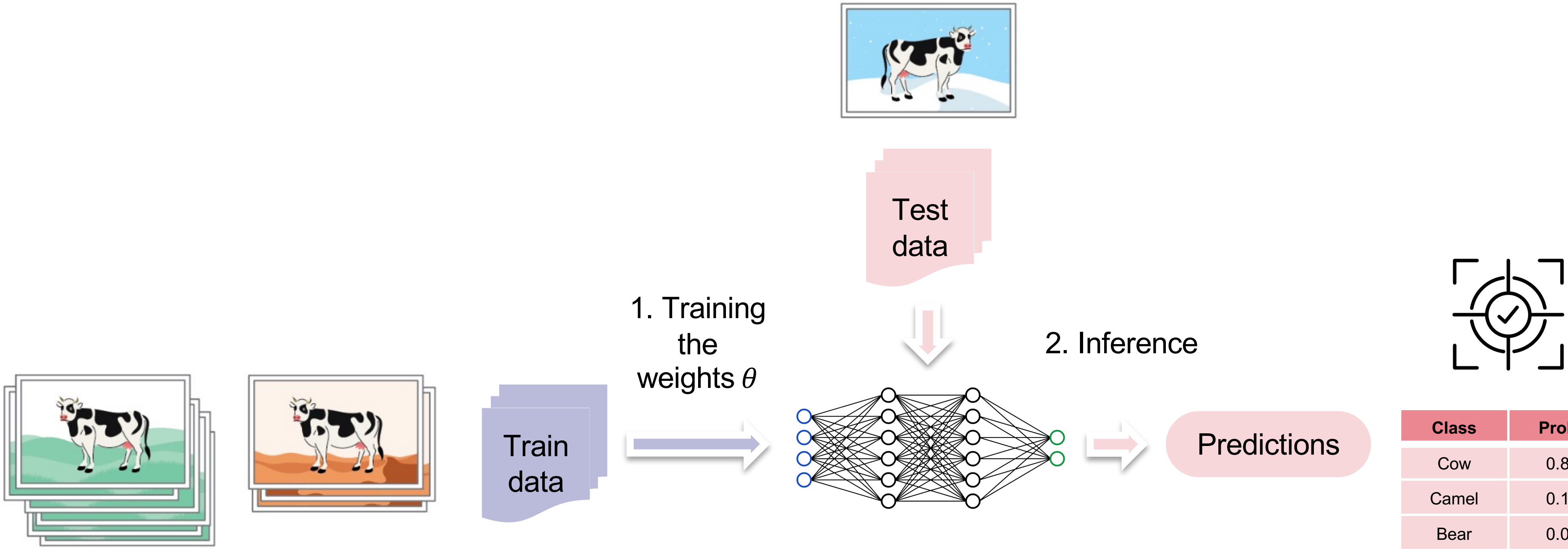
Deep learning



Ensembling in deep learning



Train and test in deep learning



Class	Proba
Cow	0.80
Camel	0.19
Bear	0.01

Out-of-distribution generalization

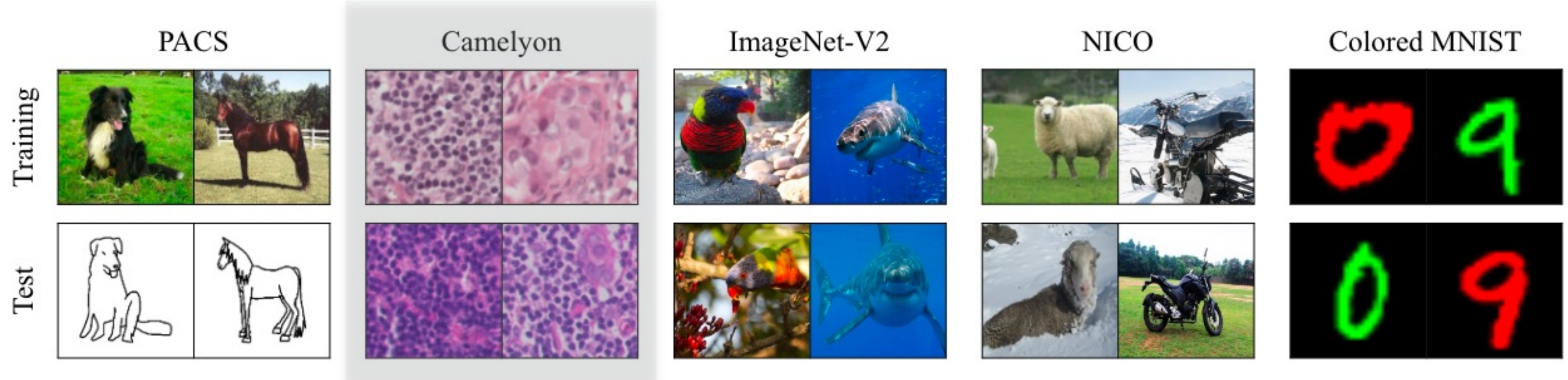
Challenge

Generalization under distribution shift, for adaptation to new domains.

Distribution shift:
$P_{train}(X, Y) \neq P_{test}(X, Y)$

Importance

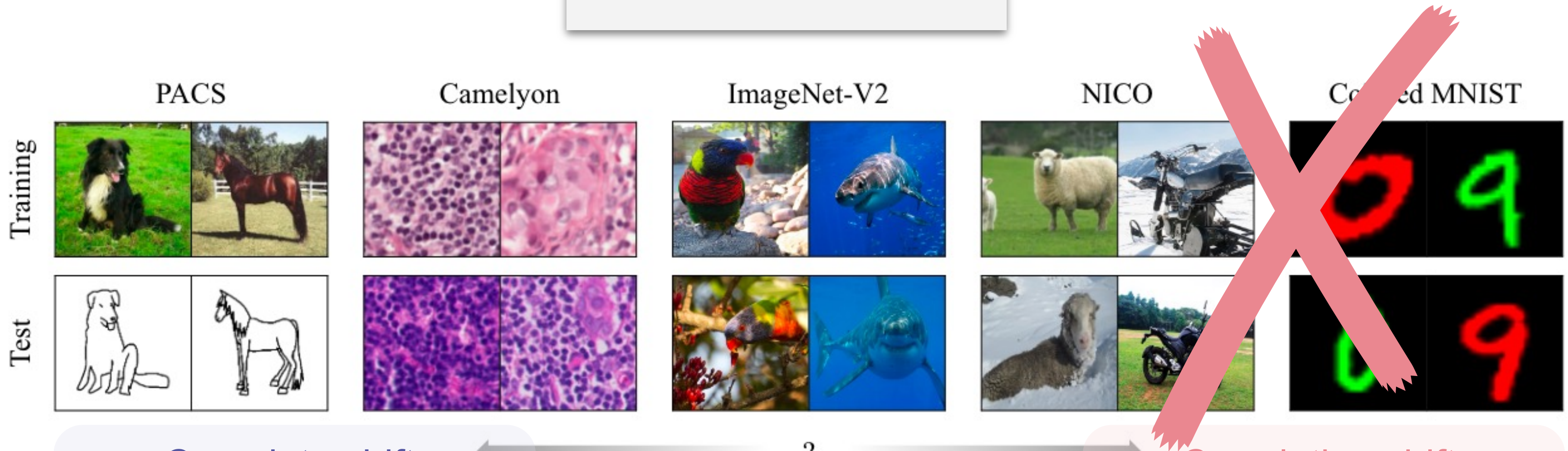
Robustness required for responsible and fair usage in most applications.



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

Two different kinds of distribution shifts

Distribution shift:
$P_{train}(X, Y) \neq P_{test}(X, Y)$



Covariate shift:
$P_{train}(X) \neq P_{test}(X)$

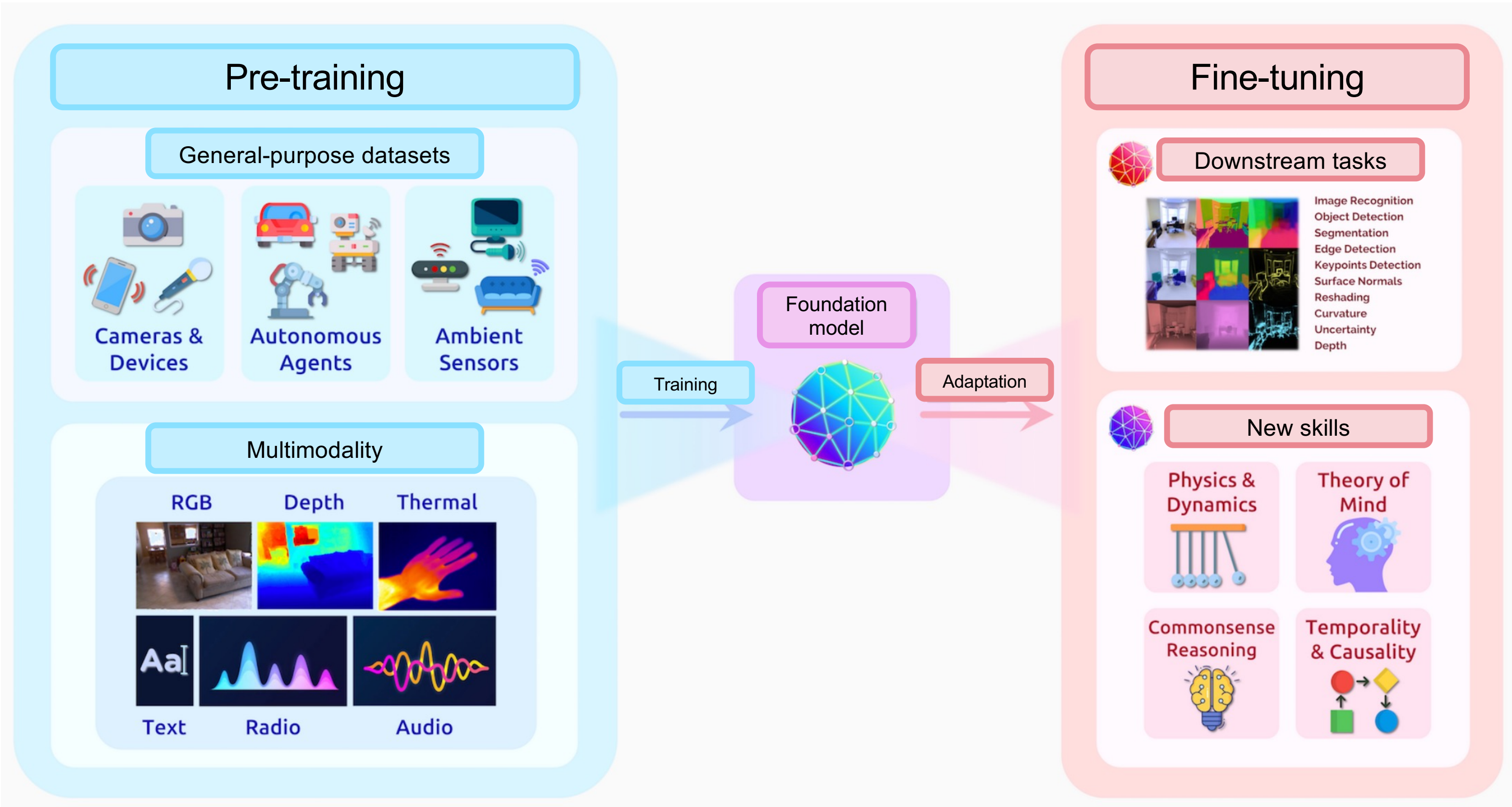
?

Correlation shift:
$P_{train}(Y|X) \neq P_{test}(Y|X)$

Fishr: Invariant Gradient Variances for Out-of-distribution Generalization.
Alexandre Ramé, Corentin Dancette, Matthieu Cord. ICML 2022.

To tackle correlation shift, please see Fishr and the invariance literature.

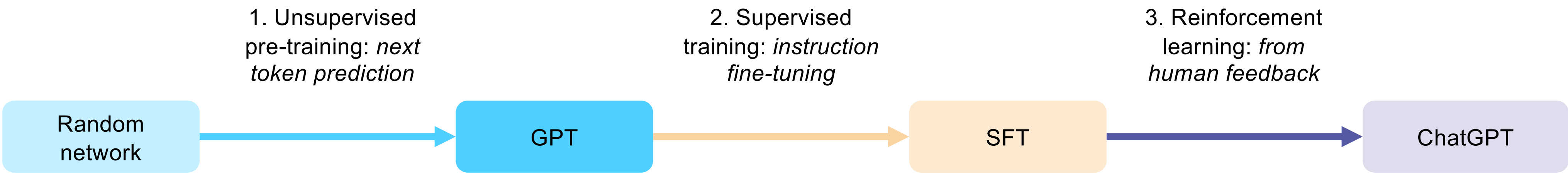
Standard strategy: fine-tuning from foundation model



[Bommasani2021] On the opportunities and risks of foundation models.

Fine-tuning and distribution shifts are everywhere

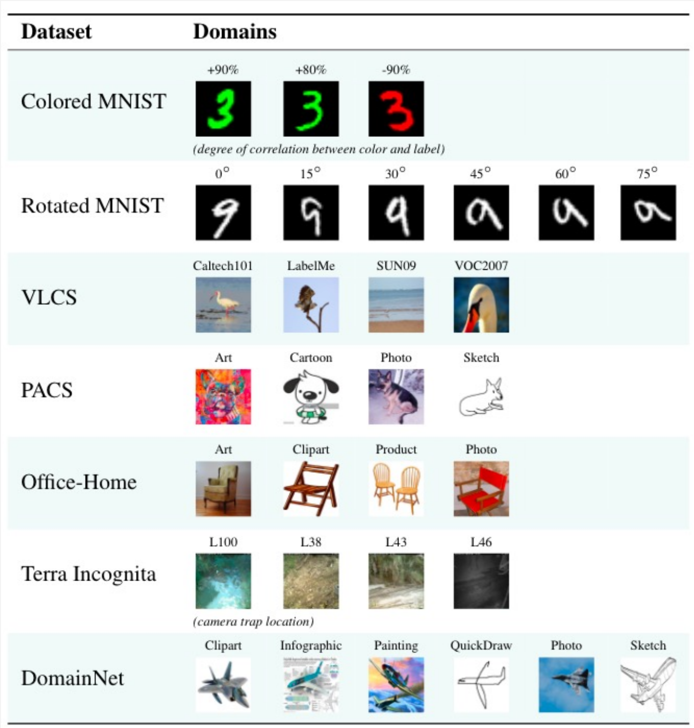
ChatGPT requires succesive fine-tunings



Fine-tunings come with out-of-distribution challenges,
even in ChatGPT.

[Ouyang2022] Training language models to follow instructions with human feedback. NeurIPS.

How to best fine-tune foundation models ?



Available algorithms

The currently available algorithms are:

- Empirical Risk Minimization (ERM, Vapnik, 1998)
- Invariant Risk Minimization (IRM, Arjovsky et al., 2019)
- Group Distributionally Robust Optimization (GroupDRO, Sagawa et al., 2020)
- Interdomain Mixup (Mixup, Yan et al., 2020)
- Marginal Transfer Learning (MTL, Blanchard et al., 2011-2020)
- Meta Learning Domain Generalization (MLDG, Li et al., 2017)
- Maximum Mean Discrepancy (MMD, Li et al., 2018)
- Deep CORAL (CORAL, Sun and Saenko, 2016)
- Domain Adversarial Neural Network (DANN, Ganin et al., 2015)
- Conditional Domain Adversarial Neural Network (CDANN, Li et al., 2018)
- Style Agnostic Networks (SagNet, Nam et al., 2020)
- Adaptive Risk Minimization (ARM, Zhang et al., 2020), contributed by @zhangmarvin
- Variance Risk Extrapolation (VREx, Krueger et al., 2020), contributed by @zdhNarsil
- Representation Self-Challenging (RSC, Huang et al., 2020), contributed by @SirRob1997
- Spectral Decoupling (SD, Pezeshki et al., 2020)
- Learning Explanations that are Hard to Vary (AND-Mask, Parascandolo et al., 2020)
- Out-of-Distribution Generalization with Maximal Invariant Predictor (IGA, Koyama et al., 2020)
- Gradient Matching for Domain Generalization (Fish, Shi et al., 2021)
- Self-supervised Contrastive Regularization (SelfReg, Kim et al., 2021)
- Smoothed-AND mask (SAND-mask, Shahtalebi et al., 2021)
- Invariant Gradient Variances for Out-of-distribution Generalization (Fishr, Rame et al., 2021)
- Learning Representations that Support Robust Transfer of Predictors (TRM, Xu et al., 2021)
- Invariance Principle Meets Information Bottleneck for Out-of-Distribution Generalization (IB-ERM , Ahuja et al., 2021)
- Invariance Principle Meets Information Bottleneck for Out-of-Distribution Generalization (IB-IRM, Ahuja et al., 2021)
- Optimal Representations for Covariate Shift (CAD & CondCAD, Ruan et al., 2022), contributed by @ryoungj
- Quantifying and Improving Transferability in Domain Generalization (Transfer, Zhang et al., 2021), contributed by @Gordon-Guojun-Zhang
- Invariant Causal Mechanisms through Distribution Matching (CausIRL with CORAL or MMD, Chevalley et al., 2022), contributed by @MathieuChevalley

The DomainBed benchmark compares different approaches.

⟹ standard *empirical risk minimization* remains the best,

until …

[Gulrajani2021] In Search of Lost Domain Generalization. ICLR.

Plan

Part I. DiWA

Diverse weight averaging for out-of-distribution generalization.

Part II. Ratatouille

Recycling diverse models for out-of-distribution generalization.

Part III. Rewarded soups

Towards Pareto-optimal alignment by interpolating weights fine-tuned on diverse rewards.

Part I. DiWA

Part I. DiWA

Diverse weight averaging for out-of-distribution generalization.

Part II. Ratatouille

Recycling diverse models for out-of-distribution generalization.

Part III. Rewarded soups

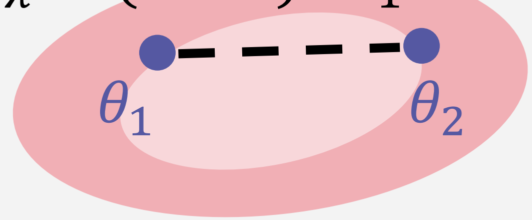
Towards Pareto-optimal alignment by interpolating weights fine-tuned on diverse rewards.

📖 DiWA: diverse weight averaging for out-of-distribution generalization.
Alexandre Ramé, Matthieu Kirchmeyer, Thibaud Rahier, Alain Rakotomamonjy, Patrick Gallinari, Matthieu Cord. NeurIPS 2022.

Weight averaging

Two weights θ_1 and θ_2 are linearly mode connected = their weight average perform well (despite the non-linearities).

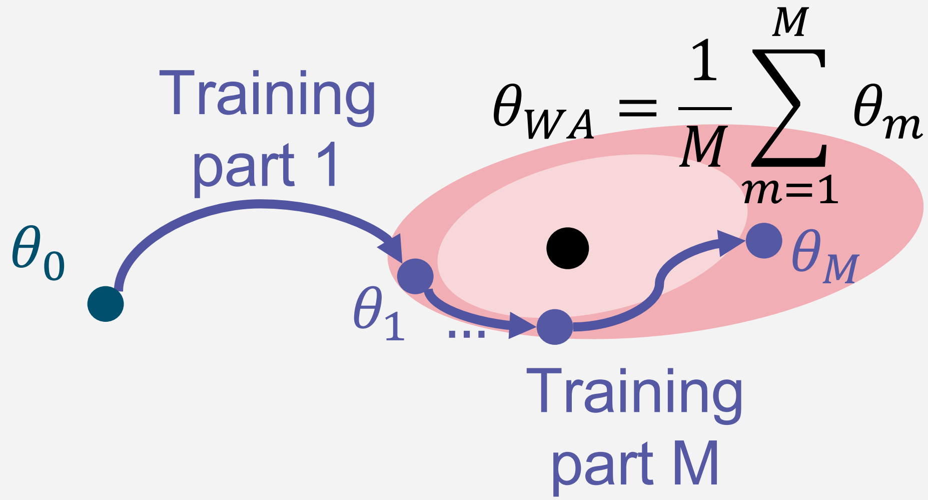
$$\theta_\lambda = (1 - \lambda) \cdot \theta_1 + \lambda \cdot \theta_2$$



Weight averaging = simple & efficient ensembling method to combine various models.

[Frankle2020] Linear mode connectivity and the lottery ticket hypothesis. ICML.

Weight averaging along a training trajectory

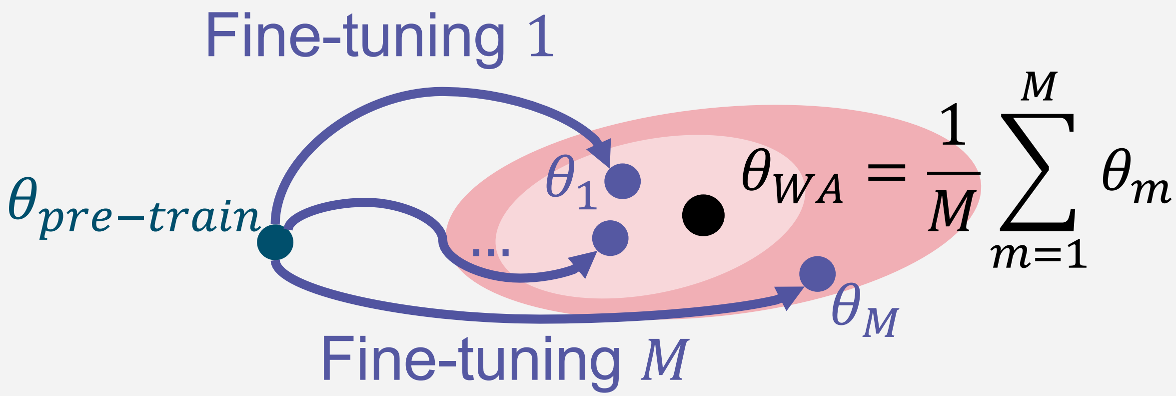


Moving average [Izmailov2018]: checkpoints collected along a training trajectory remain linearly connected.

[Izmailov2018] Averaging Weights Leads to Wider Optima and Better Generalization. UAI.

Weight averaging from multiple trajectories

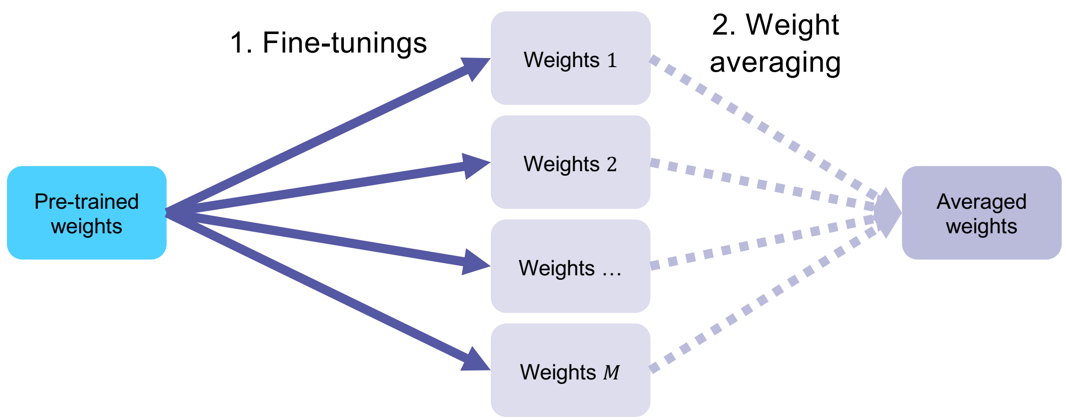
When fine-tuned from a shared pre-trained model, weights remain linearly connected.



[Neyshabur2020] What is being transferred in transfer learning? NeurIPS.

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.

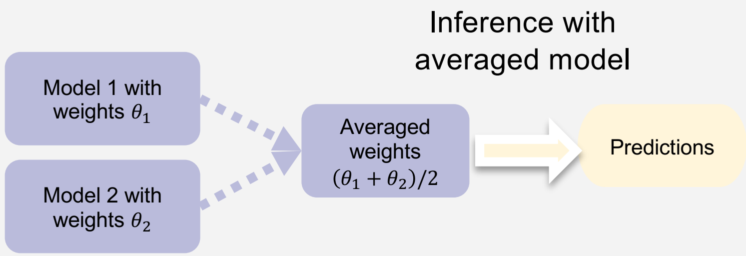
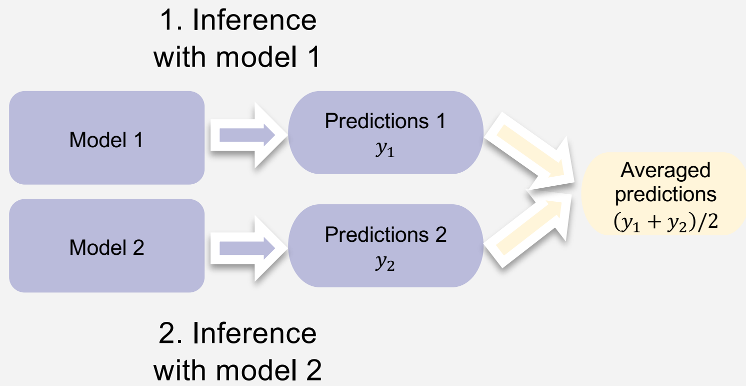
New state of the art on DomainBed

DiWA improves in-distribution generalization (state of the art on ImageNet), but gains are even more spectacular out-of-distribution.

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
DiWA	**78.4**	**88.7**	**72.1**	**51.4**	**47.4**	**67.6**

DiWA: diverse weight averaging for out-of-distribution generalization.
Alexandre Ramé, Matthieu Kirchmeyer, Thibaud Rahier, Alain Rakotomamonjy, Patrick Gallinari, Matthieu Cord. NeurIPS 2022.

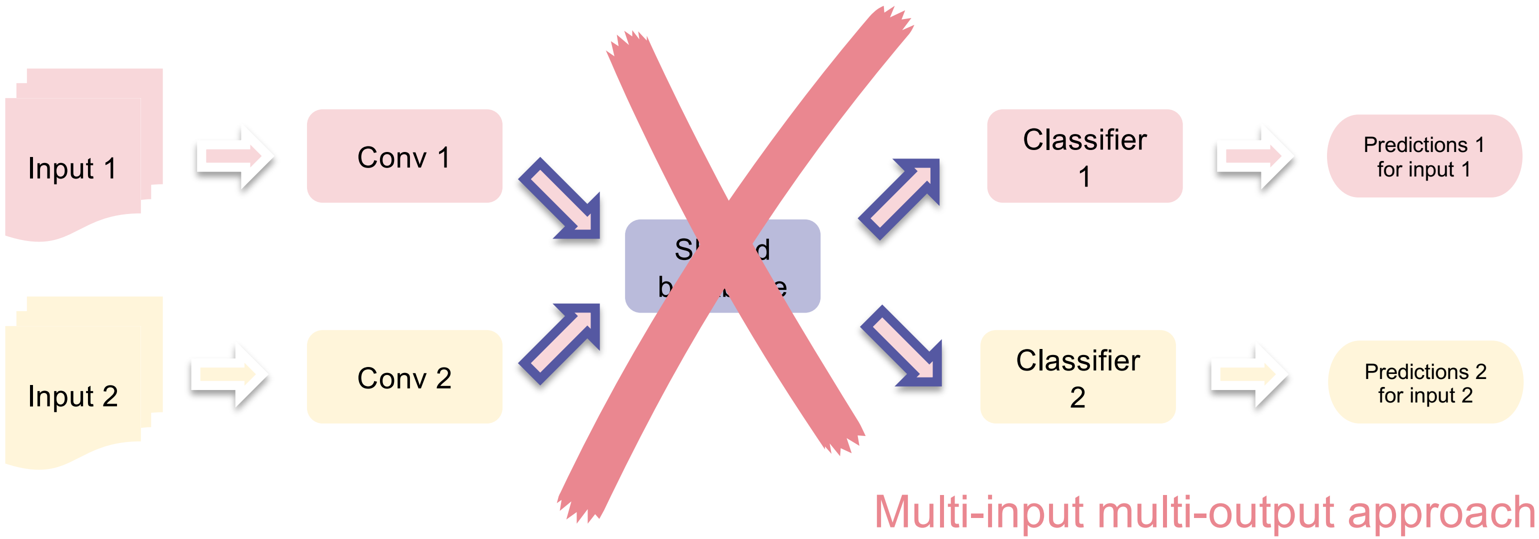
Weight averaging approximates prediction averaging

Name	Weight averaging	Prediction averaging (traditional ensembling)
What		
Inference cost	1 single forward	2 forwards
Constraint	Weights linearly mode connected for a given architecture	No constraint

More precisely, weight averaging approximates prediction averaging when $\|\theta_1 - \theta_2\|$ is small.

Efficient ensembling as a longstanding challenge

Remark: weight averaging is much simpler than other cheap ensembling methods, such as my first attempt MixMo.



Multi-input multi-output approaches
do not scale well and require training from scratch.

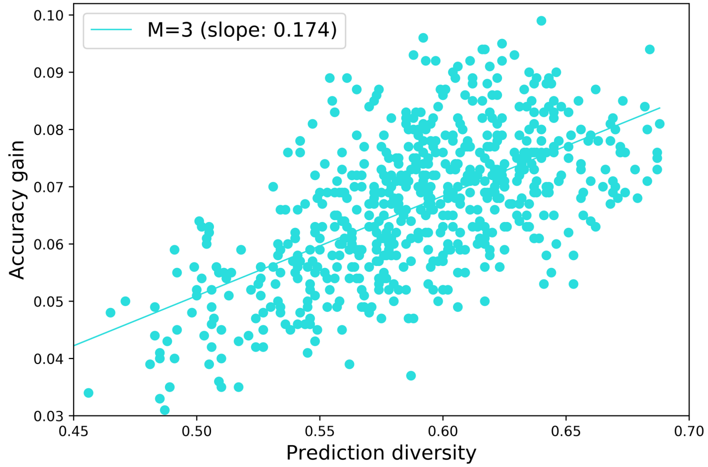
MixMo: Mixing Multiple Inputs for Multiple Outputs via Deep Subnetworks.
Alexandre Ramé, Rémy Sun, Matthieu Cord. ICCV 2021.

Diversity across averaged models improves results

Error decomposition when averaging M weights:

$$\mathbb{E}err_T(\theta_{WA}) \approx bias + \frac{1}{M} variance + \frac{M-1}{M} covariance,$$

where $covariance$ is low when models are **diverse**.



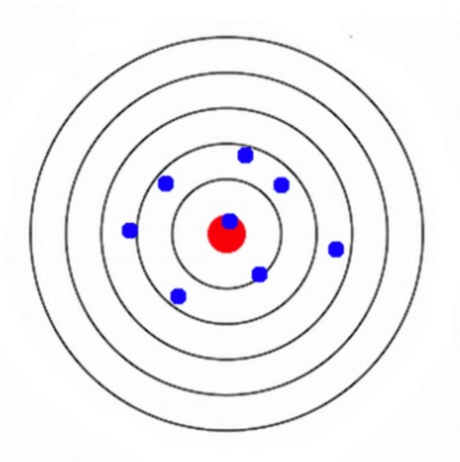
WA accuracy gain correlated with models‘ diversity.

Thus, when models are fine-tuned independently
⇒ models are more functionally **diverse**,
⇒ $covariance$ is smaller,
⇒ $\mathbb{E}err_T(\theta_{\text{DiWA}})$ is smaller than $\mathbb{E}err_T(\theta_{\text{MovingAvg}})$,
⇒ DiWA beats moving average.

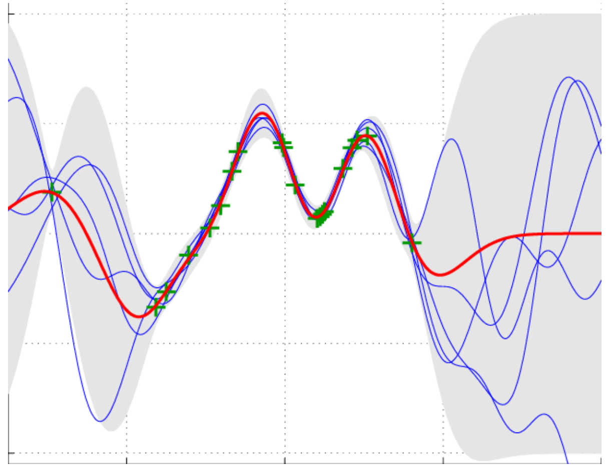
DiWA: diverse weight averaging for out-of-distribution generalization.
Alexandre Ramé, Matthieu Kirchmeyer, Thibaud Rahier, Alain Rakotomamonjy, Patrick Gallinari, Matthieu Cord. NeurIPS 2022.

Variance dominates under diversity shifts

We prove that, in the Neural Tangent Kernel regime:
$$variance \propto \text{Distance}(X_{train}, X_{test})$$



Variance is the key issue in OOD.



Networks are less constrained away from training samples.

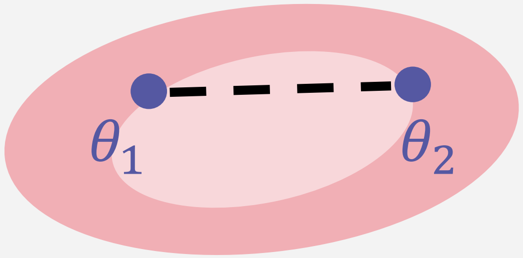
Explains why variance reduction methods
(*e.g.*, ensembling and weight averaging) are so useful in OOD.

DiWA: diverse weight averaging for out-of-distribution generalization.
Alexandre Ramé, Matthieu Kirchmeyer, Thibaud Rahier, Alain Rakotomamonjy, Patrick Gallinari, Matthieu Cord. NeurIPS 2022.

The 3 criteria for successful weight averaging

Linear connectivity

The weights should remain linearly connected.



Individual accuracies

The weights should be individually accurate.

θ_i already good when alone

Diversity

The predictions should be sufficiently diverse.

$\theta_1 \neq \theta_2$

The 3 sources of diversity in DiWA

Hyper-parameters

Sampled from a mild range.

- learning rate.
- weight decay.
- etc.

Data

- Batch orders.
- Data augmentation.
- Bagging and subsampling.

Random factors

- Dropout.
- Learning stochasticity.

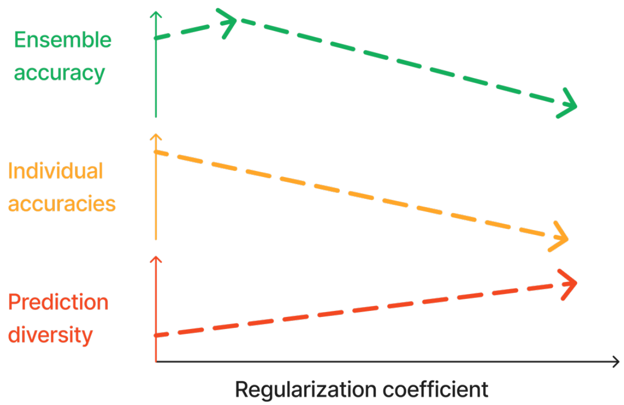
Explicit diversity ?

Idea

Regularization to increase diversity explicitly during training.

For example, the DICE information bottleneck regularization:

$$DICE = I[f_{\theta_1}(X), f_{\theta_2}(X)|Y]$$



Limitation

Complex implementation because *individual accuracies are reduced when increasing diversity with large regularization coefficient*.

DICE: Diversity in Deep Ensembles via Conditional Redundancy Adversarial Estimation. **Alexandre Ramé** and Matthieu Cord. ICLR 2021.

Diversity in initialization?

Idea

Different initializations to increase diversity.

Problem

Weights with different initializations are not linearly connected.

First strategy

Permutation to align weights [Ainsworth2023].
⇒ poor results empirically.

[Ainsworth2023] Git Re-Basin: Merging Models modulo Permutation Symmetries. ICLR.

Part II. Ratatouille

Part I. DiWA

Diverse weight averaging for out-of-distribution generalization.

Part II. Ratatouille

Recycling diverse models for out-of-distribution generalization.

Part III. Rewarded soups

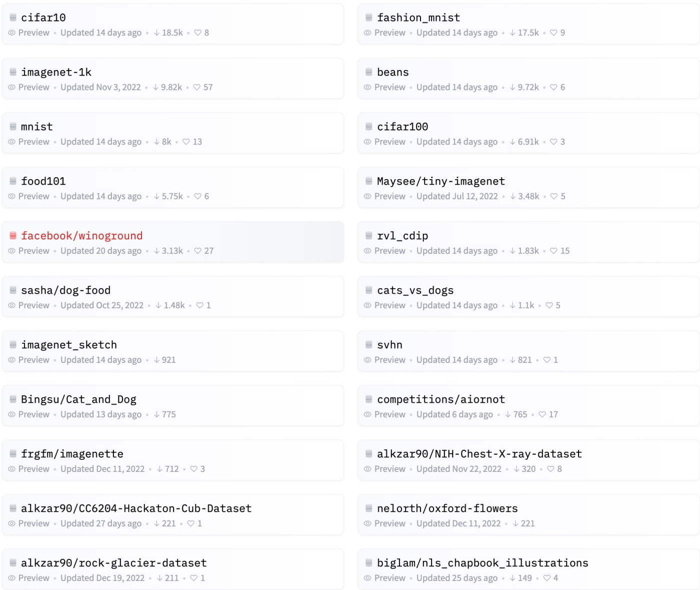
Towards Pareto-optimal alignment by interpolating weights fine-tuned on diverse rewards.

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.

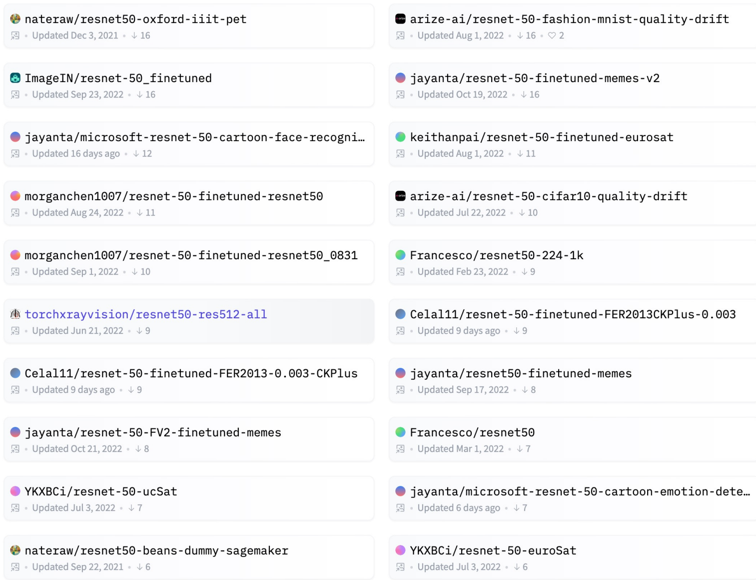
Meta

Era of open-source datasets and weights

huggingface.co/datasets



huggingface.co/models/resnet-50

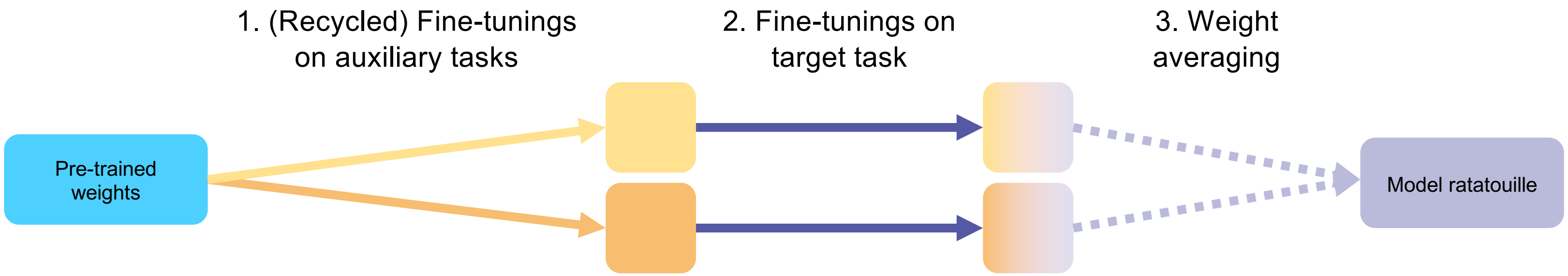


Key idea: recycle these weights as initializations for the target task.

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.

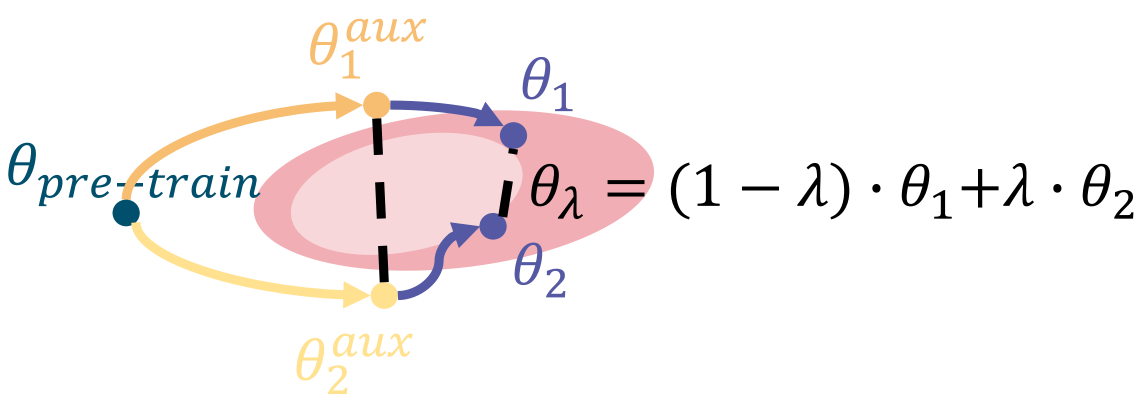


Model ratatouille: recycling diverse models for out-of-distribution generalization.
Alexandre Ramé, Kartik Ahuja, Jianyu Zhang, Matthieu Cord, Léon Bottou and David Lopez-Paz. ICML 2023.

Does Ratatouille meet the 3 criteria for weight averaging ?

Linear connectivity

Yes, the weights remain linearly connected when auxiliary tasks are sufficiently similar.



Does Ratatouille meet the 3 criteria for weight averaging ?

Linear connectivity

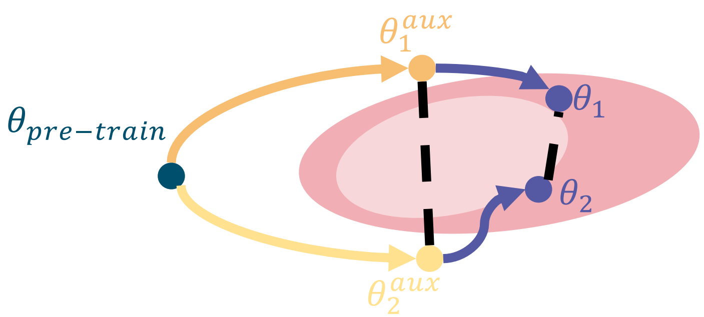
Yes, the weights remain linearly connected when auxiliary tasks are sufficiently similar.

Individual accuracies

Yes, when the auxiliary tasks learn rich features, that help for the target task.

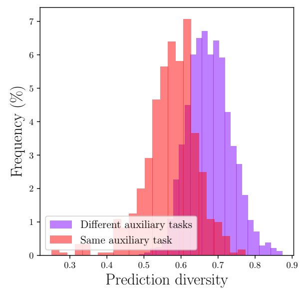
Diversity

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

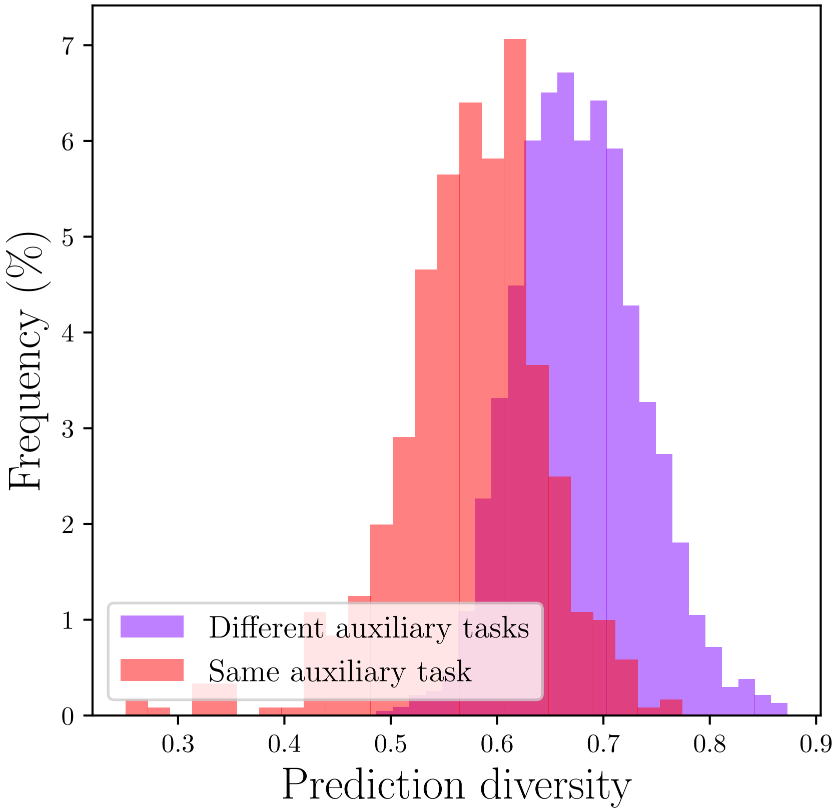


[Phang2018] Sentence encoders on stilts: Supplementary training on intermediate labeleddata tasks.

[Choshen2022] Where to start? analyzing the potential value of intermediate models.



Inter-trainings on different auxiliary tasks increase diversity



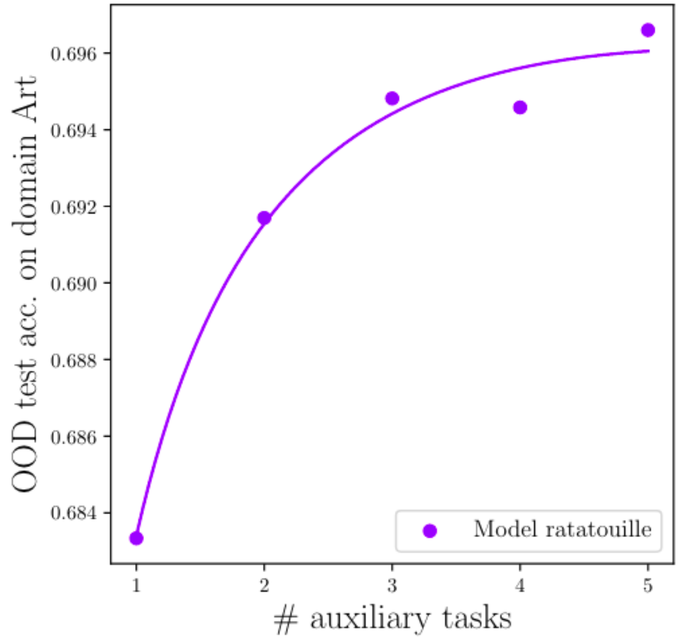
Diversity

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

New (improved) state of the art on DomainBed

- No training overhead (if auxiliary weights are recycled).
- Auxiliary datasets: those from 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
DiWA	78.4	88.7	72.1	51.4	47.4	67.6
Ratatouille	**78.5**	**89.5**	**73.1**	**51.8**	**47.5**	**68.1**



Part III. Rewarded soups

Part I. DiWA

Diverse weight averaging for out-of-distribution generalization.

Part II. Ratatouille

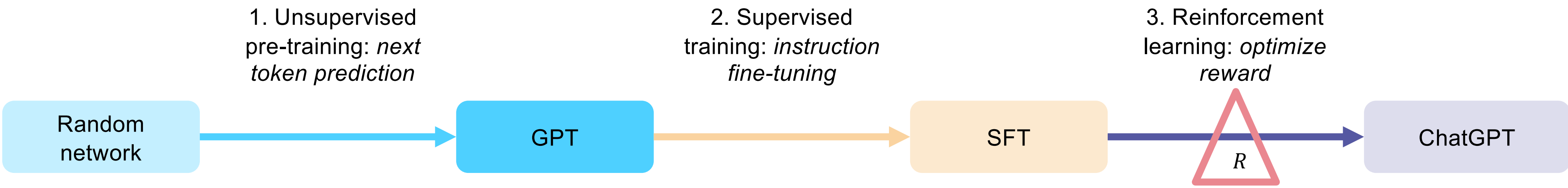
Recycling diverse models for out-of-distribution generalization.

Part III. Rewarded soups

Towards Pareto-optimal alignment by interpolating weights fine-tuned on diverse rewards.

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.

Training strategy for large language models

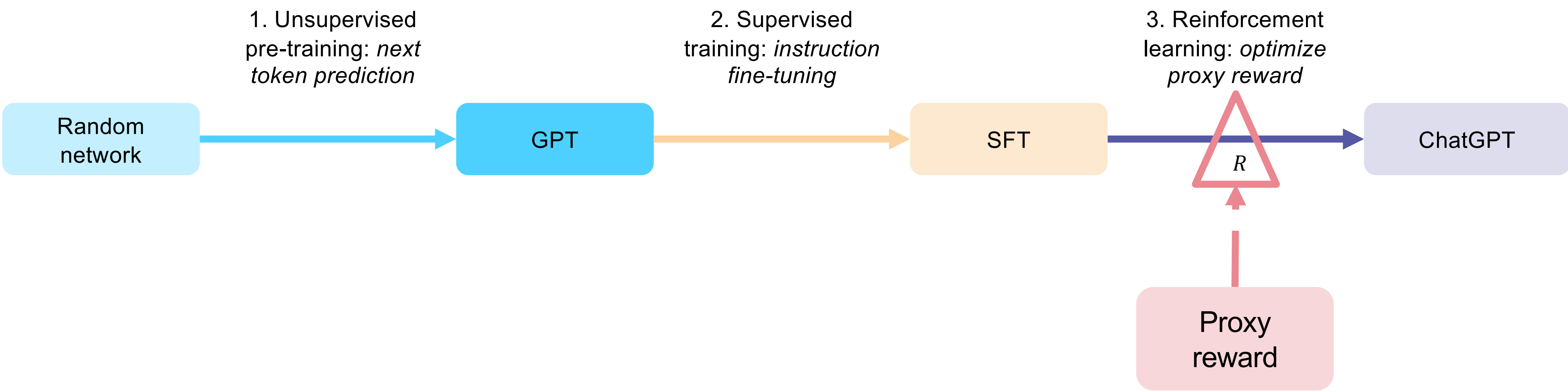


Why RL:

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

[Ouyang2022] Training language models to follow instructions with human feedback. NeurIPS.

The risk of reward misspecification



Problem: the true reward is not available.
Solution: **proxy reward**.
Challenge: designing **reliable** proxy rewards is hard.
Risk: reward **misspecification**.

[Ouyang2022] Training language models to follow instructions with human feedback. NeurIPS.

Reinforcement learning from human feedback (RLHF)

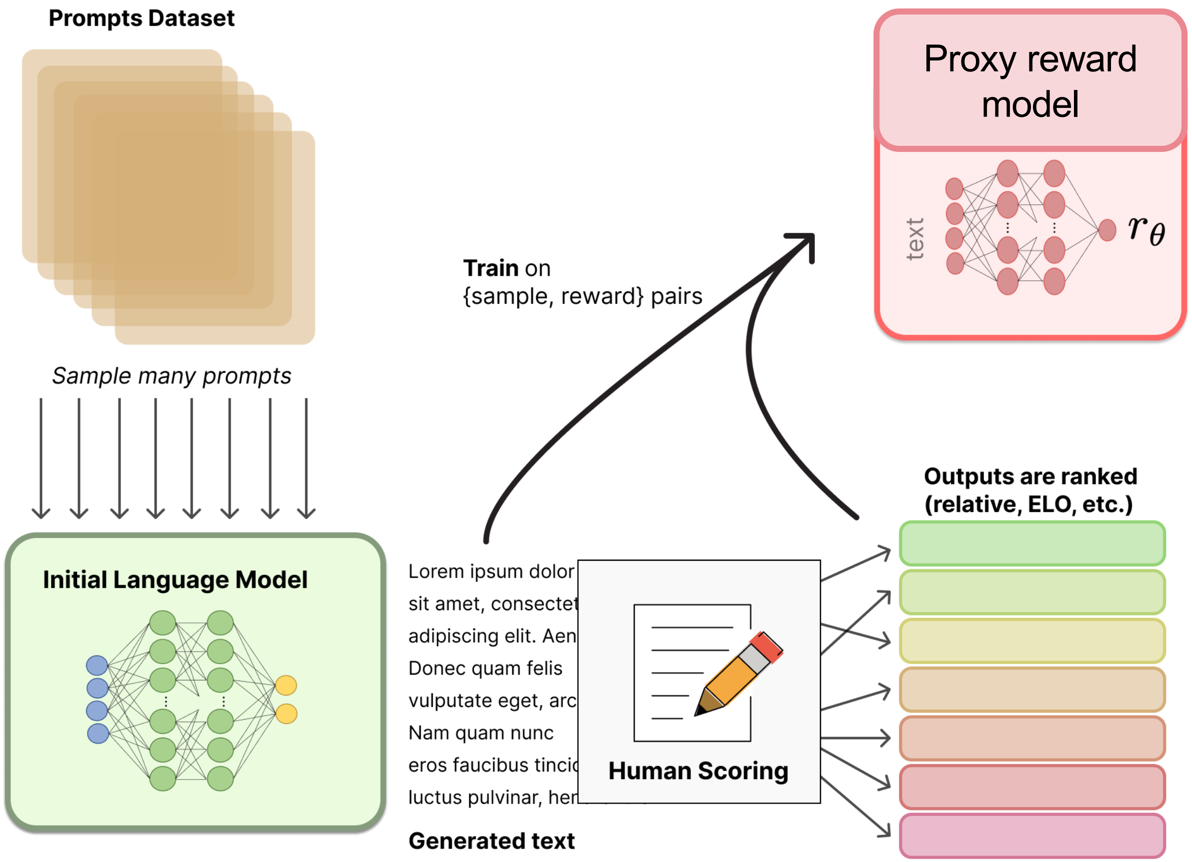


Image from https://huggingface.co/blog/rlhf.

Diversity of opinions

Consistency issue: only $\approx 65\%$ agreement across labellers.

Humans have **diverse opinions** (politics, aesthetics, etc) and **different expectations** from machines (helpfulness *vs.* harmlessness).

Diversity of opinions ⇒ which one should we optimize for?

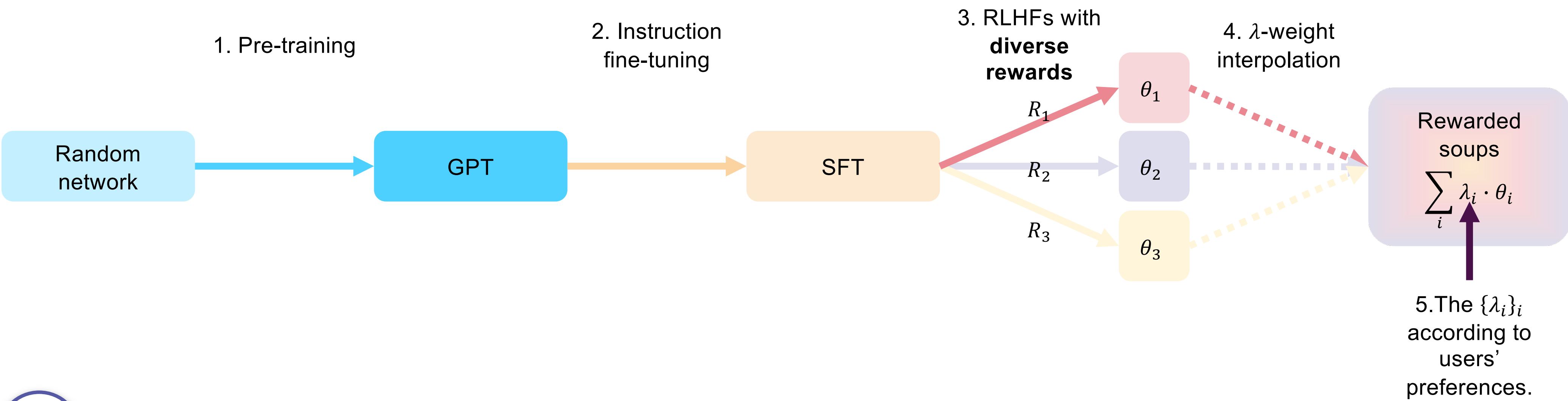
Embrace the diversity of human opinions

From a single-policy towards a **multi-policy** paradigm:

“Human aligned artificial intelligence is a multi-objective problem” [Vamplew2018].

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).



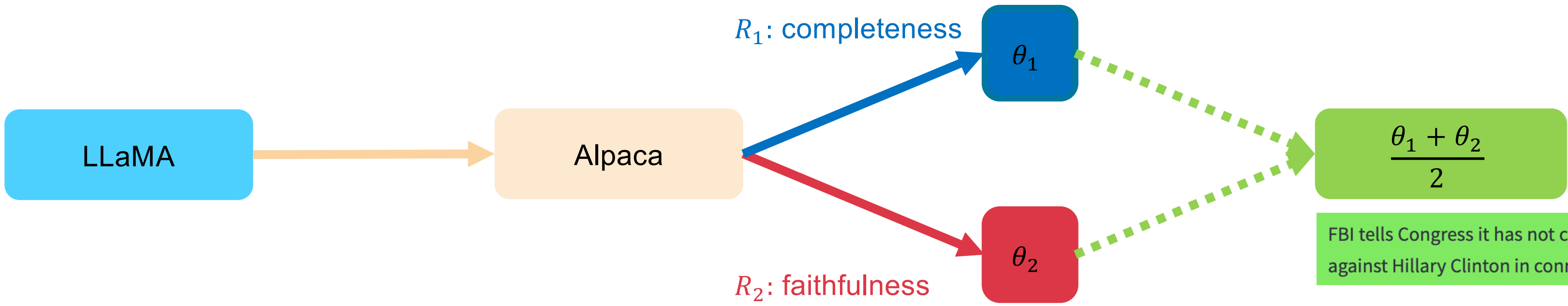
Rewarded soups: towards Pareto-optimal alignment by interpolating weights fine-tuned on diverse rewards.
Alexandre Ramé, Guillaume Couairon, Corentin Dancette, Jean-Baptiste Gaya, Mustafa Shukor, Laure Soulier, Matthieu Cord. NeurIPS 2023.

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.

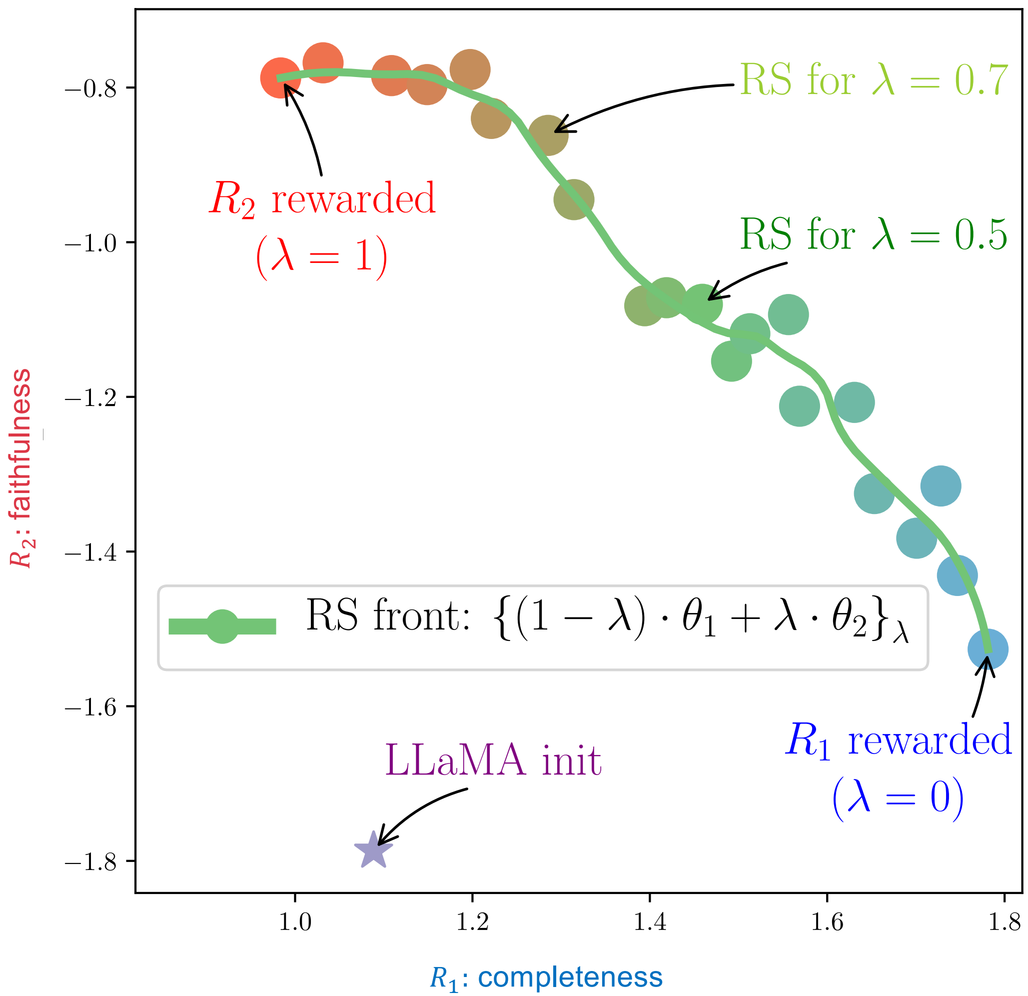


FBI tells Congress it has not changed its conclusion that no charges should be brought against Hillary Clinton in connection with her use of a private email server.

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

[Touvron2023] LLaMA: Open and efficient foundation language models.
[Taori2023] Stanford Alpaca: An instruction-following LLaMA model.

Summarization: completeness *vs.* faithfulness



Hillary Clinton email controversy

Text to summarize:

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.

Generated summaries:

λ=0.0

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.

λ=0.7

FBI tells Congress it has not changed its original decision not to pursue charges against Hillary Clinton in the email scandal.

λ=1.0

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

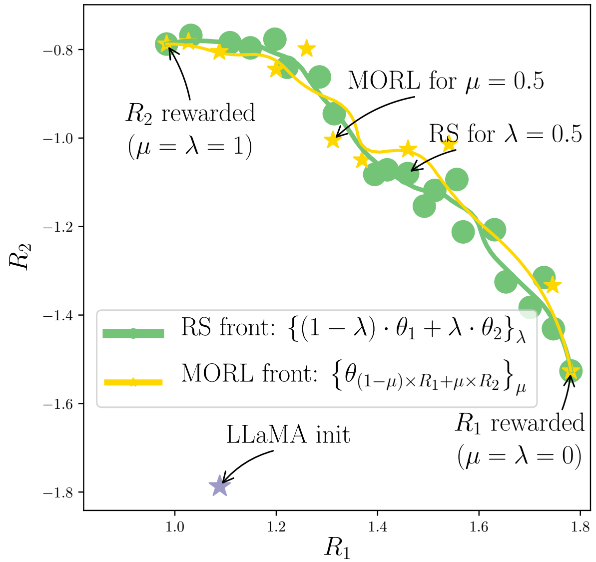
Pareto-optimal alignment across rewards

Rewarded soups

Interpolate the weights a posteriori:

$$\sum_i \lambda_i \cdot \theta_i$$

In the paper, we theoretically prove the (approximated) Pareto-optimality of rewarded soups for quadratic rewards.



MORL

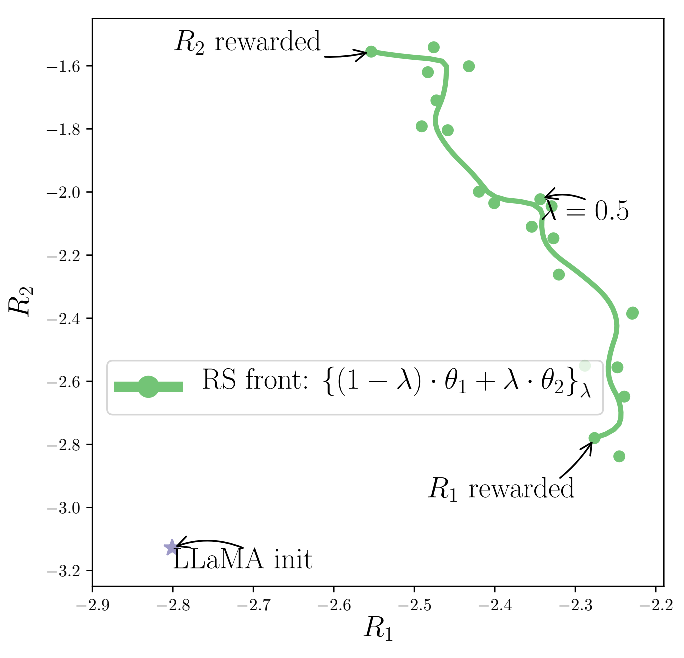
Multi-objective RL interpolate the rewards a priori:

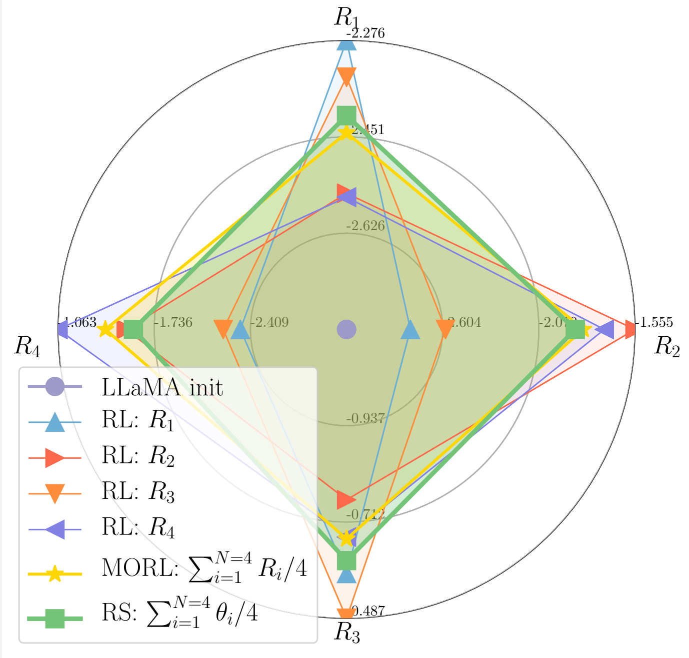
$$\sum_i \mu_i \cdot R_i$$

Issue: cost, as preference variations result in different solutions, requiring a high level of granularity.

LLaMA for conversational assistant

- Task: conversational assistant.
- Model: LLaMA-7b + Alpaca.
- Rewards: 4 OpenAssistant rewards from HuggingFace.





[Köpf2023] OpenAssistant Conversations - Democratizing Large Language Model Alignment.

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

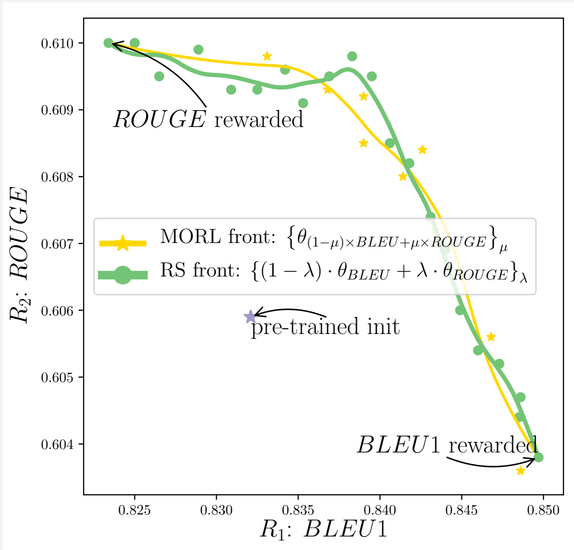
- Robot continuous control.

Captioning with diverse statistical rewards

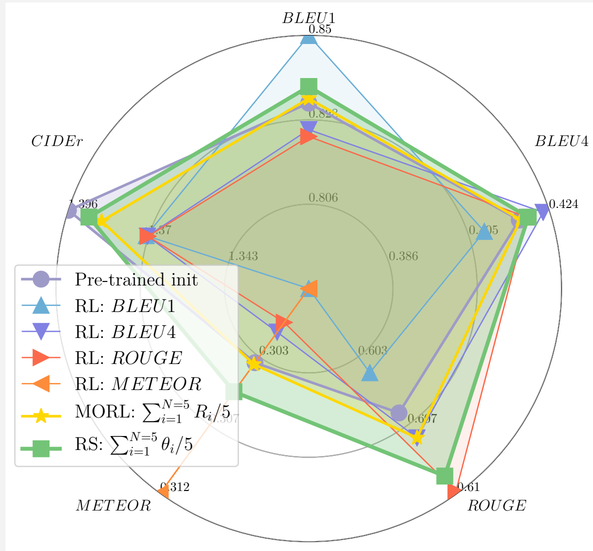
- Task: describe an image.
- Model: ExpansionNet v2 state-of-the-art initialization.
- Rewards: hand-engineered metrics:
  - The precision-focused BLEU,
  - The recall-focused ROUGE,
  - METEOR handling synonyms,
  - CIDEr using tf-idf.

Ours: A man riding a wave in the ocean.

GT: A man riding a wave on a surfboard in the ocean.



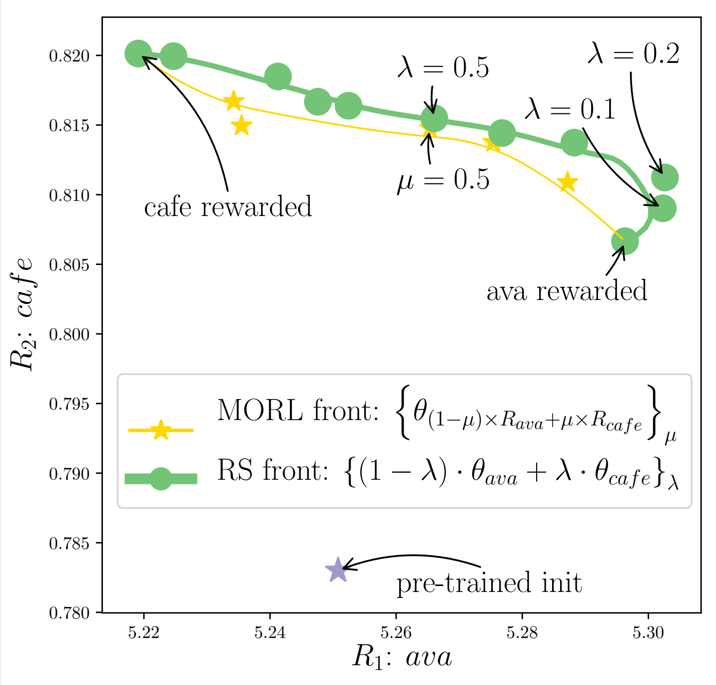
BLEU1 and ROUGE



Spider map, uniform averaging

Image generation with diverse RLHFs

- Task: align text-to-image generation with human feedback.
- Model: diffusion model with 2.2B parameters (same quality as Stable Diffusion).
- Reward: *ava* and *café* aesthetic models.



Benefits from rewarded soups

Efficiency

- 1 fine-tuning per reward.
- Parallelizable
- No inference overhead.
- Iterative and continual alignment by updating λ.

Transparence

- Support decision-making.
- Facilitate regulation by non-technical committee.
- Less engineering choices.

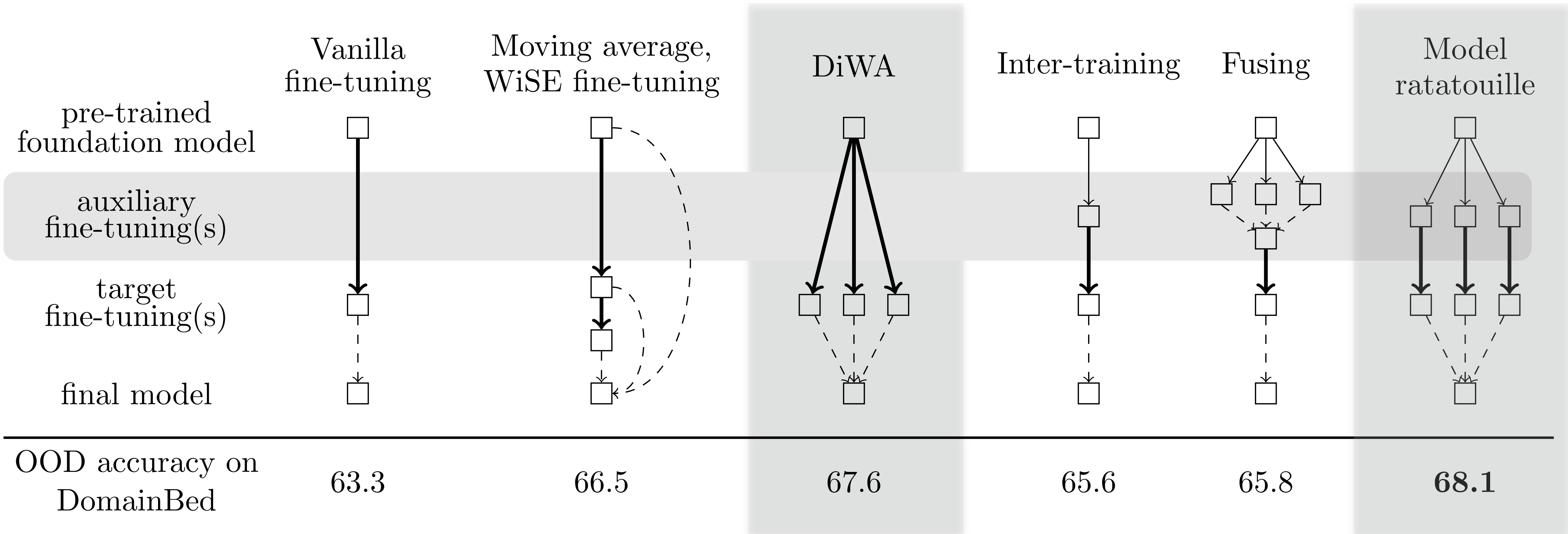
Fairness

- Value pluralism.
- Tailored for minorities.
- Less ideological hegemony.

Conclusion

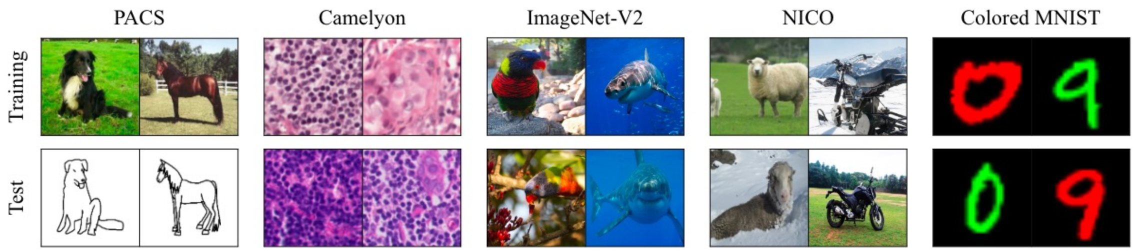
Summary of contributions and perspectives

1st contribution: improved fine-tuning strategies



2nd contribution: **diversity** for robust ensembling

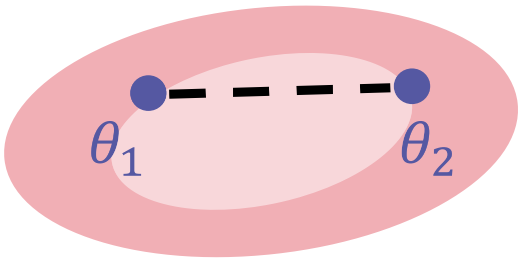
Combining diverse members as a general–purpose robustness strategy to handle train-test differences.
- distribution shifts for out-of-distribution generalization
- reward misspecification for alignment



3^{rd} contribution: weight averaging for **efficient** ensembling

Linear mode connectivity in all considered scenarios:
- various setups: supervised and reinforcement learning.
- various tasks: classification or generation.
- various modalities: text and image.

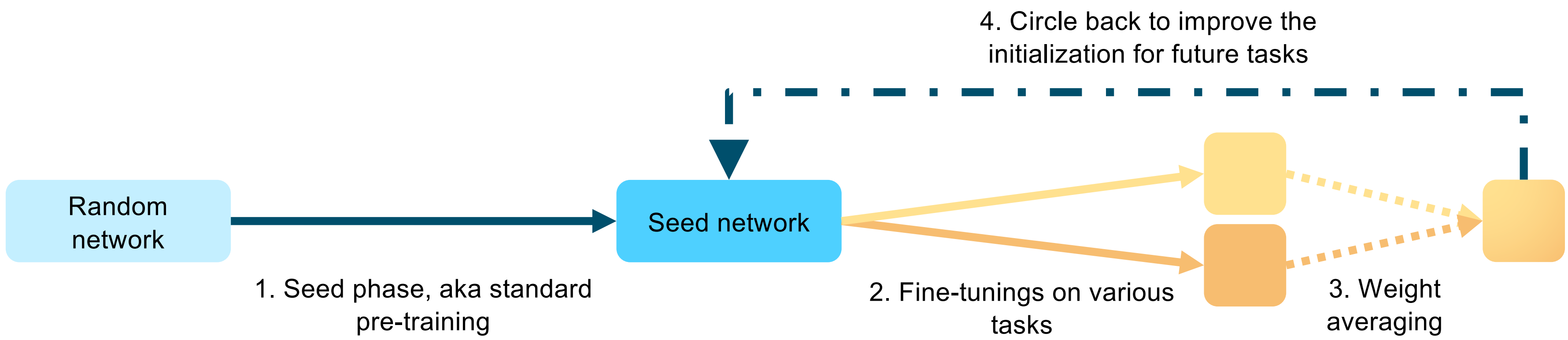
And thus weight averaging as a **scalable** strategy for foundation models.



The larger the model,
the easier the weight averaging.

4^{th} contribution: framework for **large-scale** training

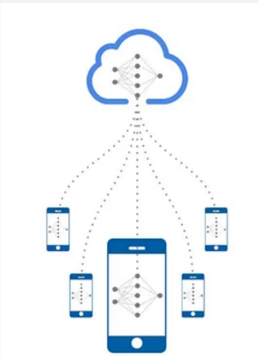
Promote a scalable training framework extending the foundation paradigm:
1. pre-training of foundation models.
2. parralelizable fine-tunings on various tasks.
3. weight averaging to combine information.
4. iterate



Towards updatable machine learning

Data privacy

Only share weights, data remain private $\Longrightarrow$ scalable federated strategy.



Compute parallelism

Embarrassingly simple parallelization with multiple independent trainings.

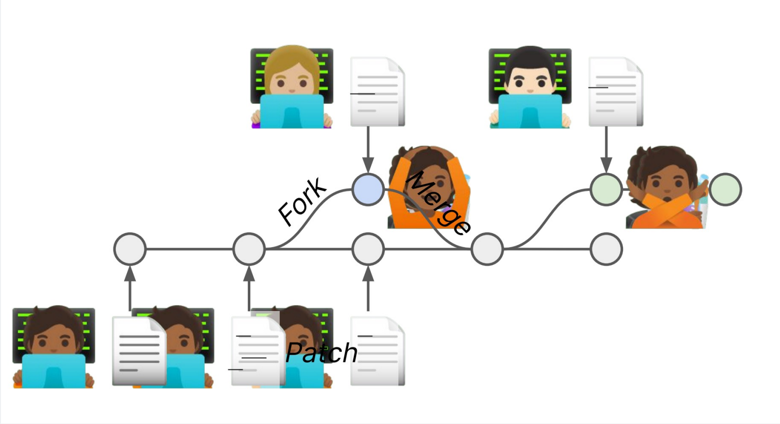
Open source

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

$\Longrightarrow$ new collaborative solutions.

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

Towards training models like we build softwares

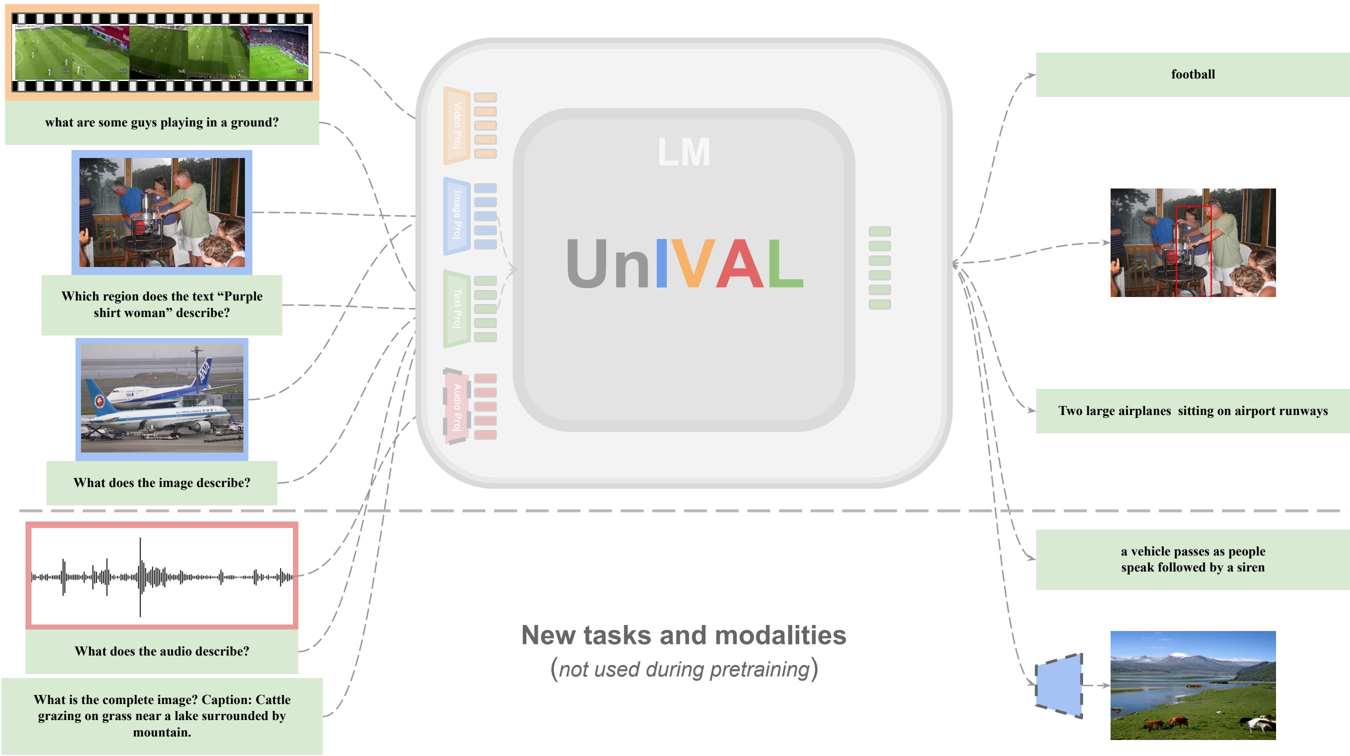


Git: software engineering	GitML: machine learning
Init	Pre-training
Commit	Fine-tuning on a task
Branch merging	Weight averaging
Unit tests	Evaluation on datasets
Merge conflict	Weight permutation (if no connectivity)

Key idea: networks as pieces of software, updatable with a GitML version control.

[Kandpal2023] Git-Theta: A Git Extension for Collaborative Development of Machine Learning Models. ICML.

Towards unified and aligned multimodal models



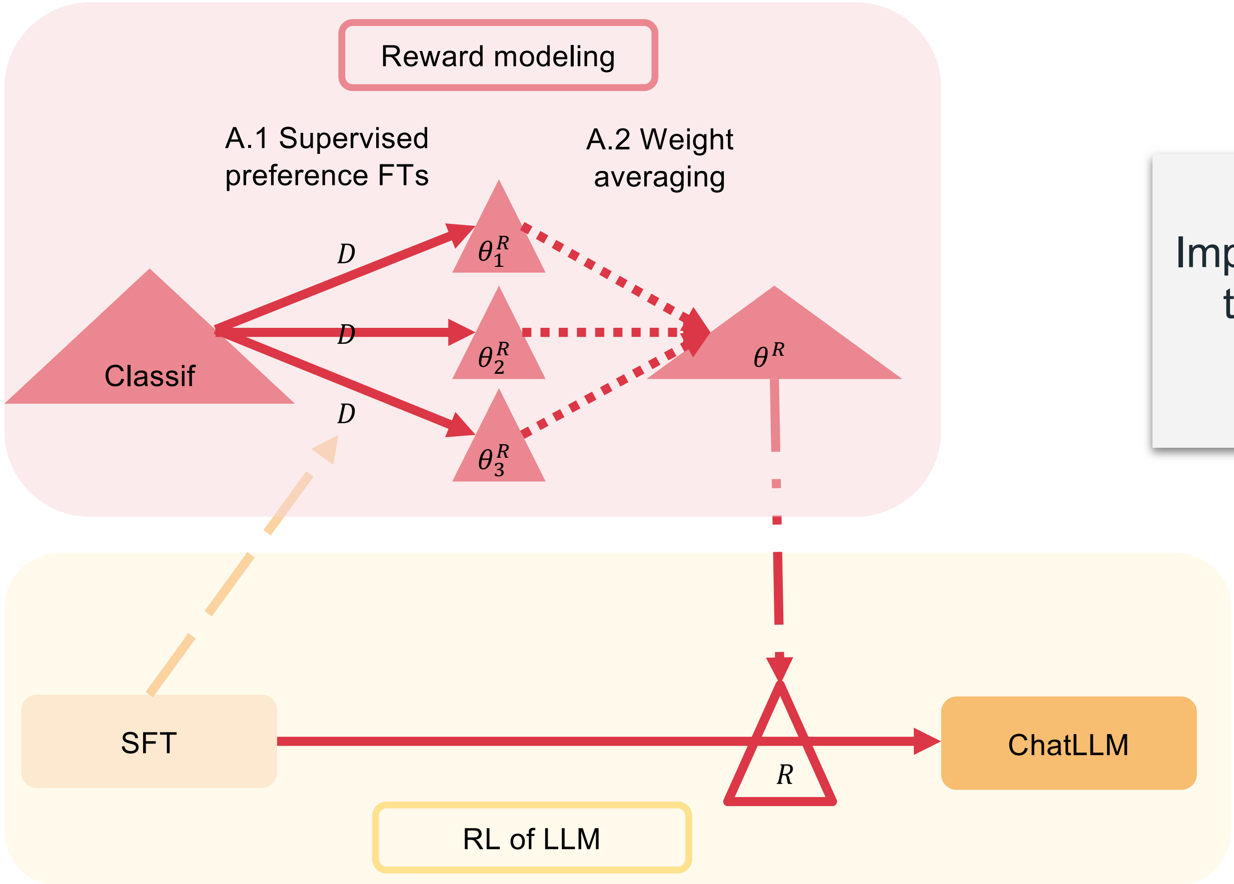
Flaws in multimodal models:
- object hallucination,
- lack of explainability,
- etc.

RLHF of multimodal models?

UniVAL: Unified Model for Image, Video, Audio and Language Tasks
Mustafa Shukor, Corentin Dancette, **Alexandre Ramé**, Matthieu Cord. 2023. In submission.

Beyond task performance: evaluating and reducing the flaws of large multimodal models with in-context learning
Mustafa Shukor, **Alexandre Ramé**, Corentin Dancette, Matthieu Cord. 2023. In submission.

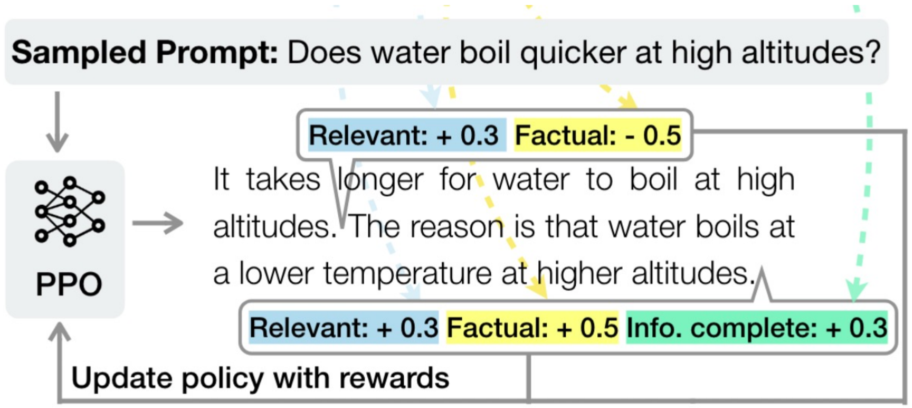
Towards more robust rewards with weight averaging

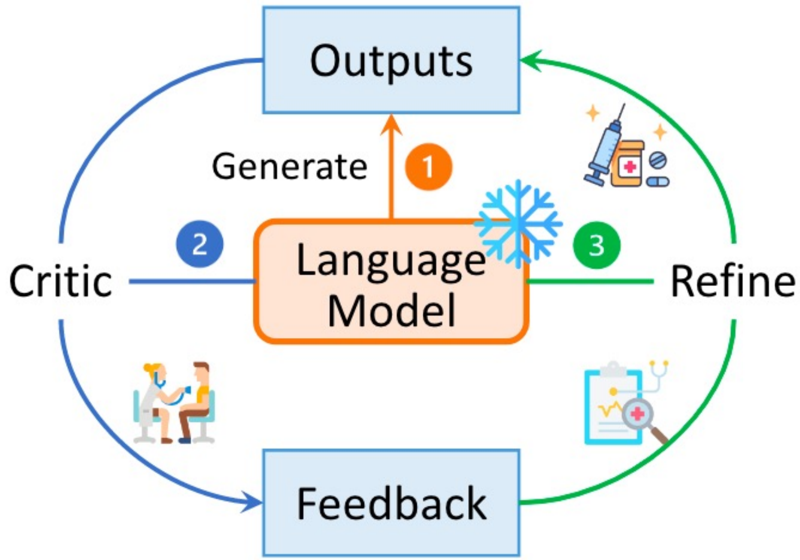


Improve the reward model with the tools from the OOD literature, such as weight averaging.

Towards fine-grained AI feedback

Problem: feedback by humans becomes inconsistent or even impossible when fine-grained.
Solution: fine-grained self-evaluation by AIs!
Consequence: towards **iterated amplification** with multiple fine-grained rewards.





[Bai2022] Constitutional AI: harmlessness from AI feedback.
[Wu2023] Fine-Grained Human Feedback Gives Better Rewards for Language Model Training. NeurIPS.
[Lee2023] RLAIF: Scaling reinforcement learning from human feedback with AI feedback.

Final perspective

How the generalization literature can help for alignment, to improve performances while mitigating ethical issues and safety risks.

Thank you for your attention

Name	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
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	Submission	2023
EvALign	Evaluating and reducing the flaws of LMMs with in-context-learning?	Submission	2023

SORBONNE UNIVERSITÉ

ISIR

MLIA Machine Learning & Deep Learning for Information Access