Rewarded soups:

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

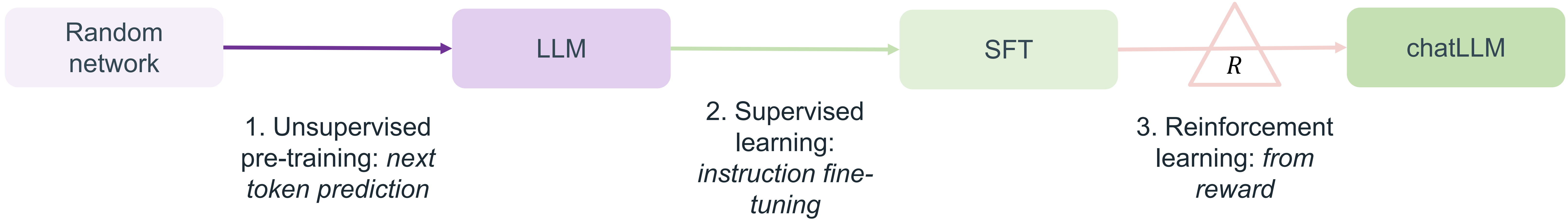
Alexandre Ramé, Guillaume Couairon, Corentin Dancette, Jean-Baptiste Gaya, Mustafa Shukor,
Laure Soulier, Matthieu Cord.

NeurIPS 2023

SCIENCES SORBONNE UNIVERSITÉ

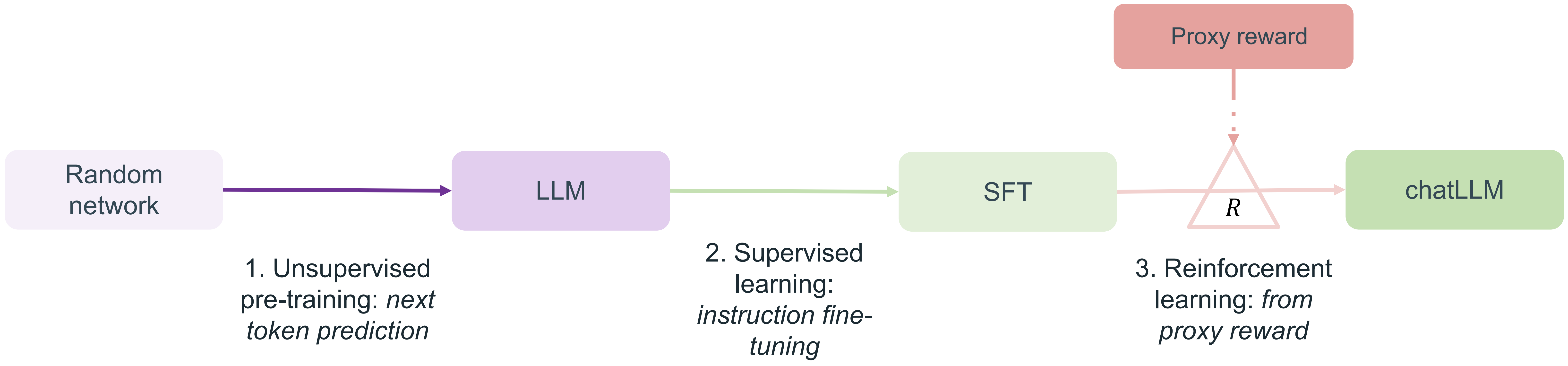
NEURAL INFORMATION PROCESSING SYSTEMS

LLM training in 3 steps



[Stiennon2020] Learning to summarize from human feedback. NeurIPS.
[Ouyang2022] Training language models to follow instructions with human feedback. NeurIPS.

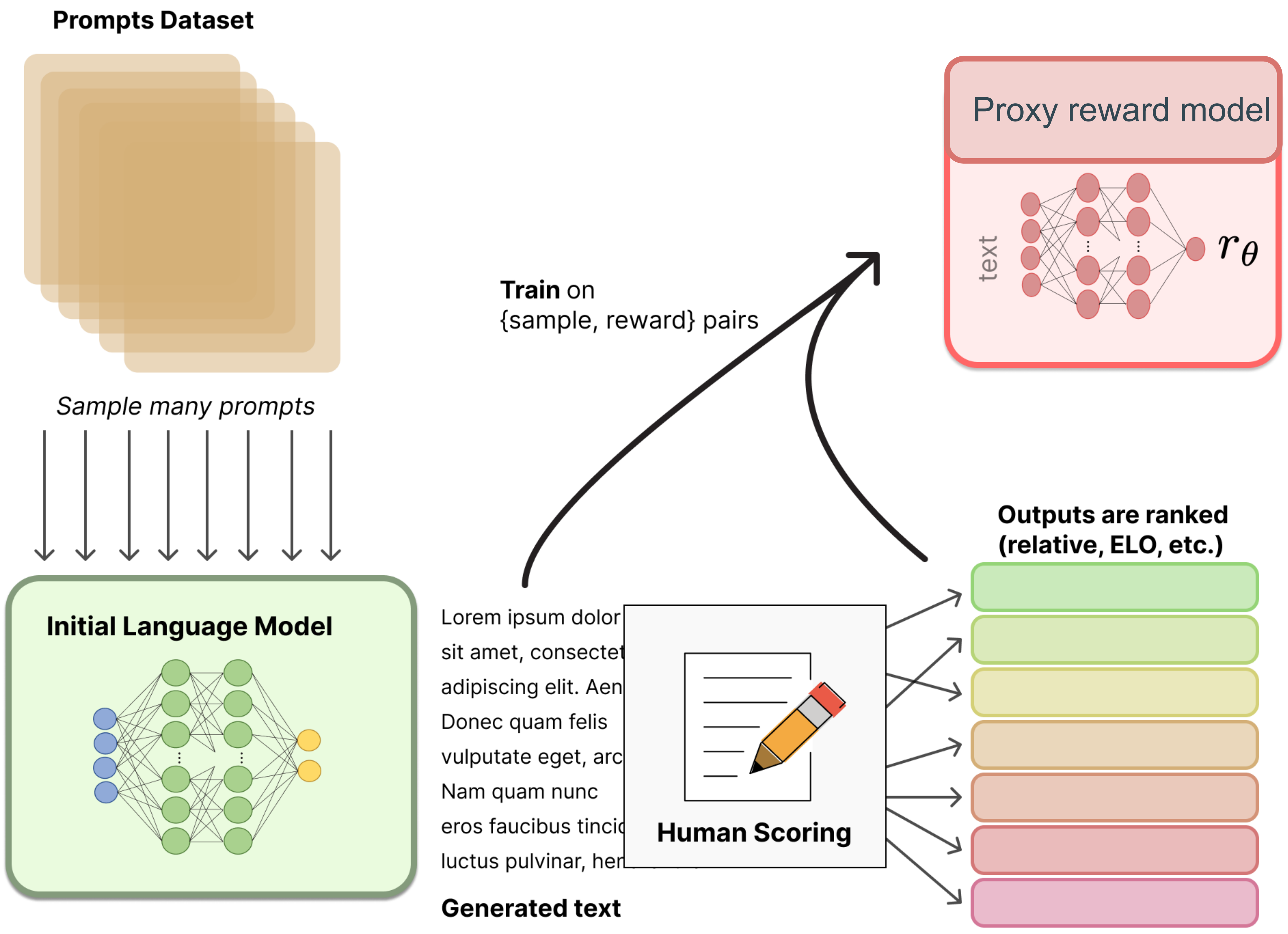
Need for a proxy reward in the RL step



Problem: the true reward is not available.
Consequence: proxy reward.
Challenge: reward misspecification
(when the training reward is not a good proxy).

[Stiennon2020] Learning to summarize from human feedback. NeurIPS.
[Ouyang2022] Training language models to follow instructions with human feedback. NeurIPS.

Reward model from human feedback for RLHF



[Christiano2017] Deep reinforcement learning from human preferences. NeurIPS.
Image from https://huggingface.co/blog/rlhf.

Diversity of opinions

Consistency issue: only ≈65% agreement across labellers.

Indeed, human opinions are diverse (and subjective):

- **Politics**: democrat vs republican?
- **Uncertain situations**: economic strategy for climate change?
- **Aesthetic**: beautiful vs ugly?

More generally, different expectations from machines:

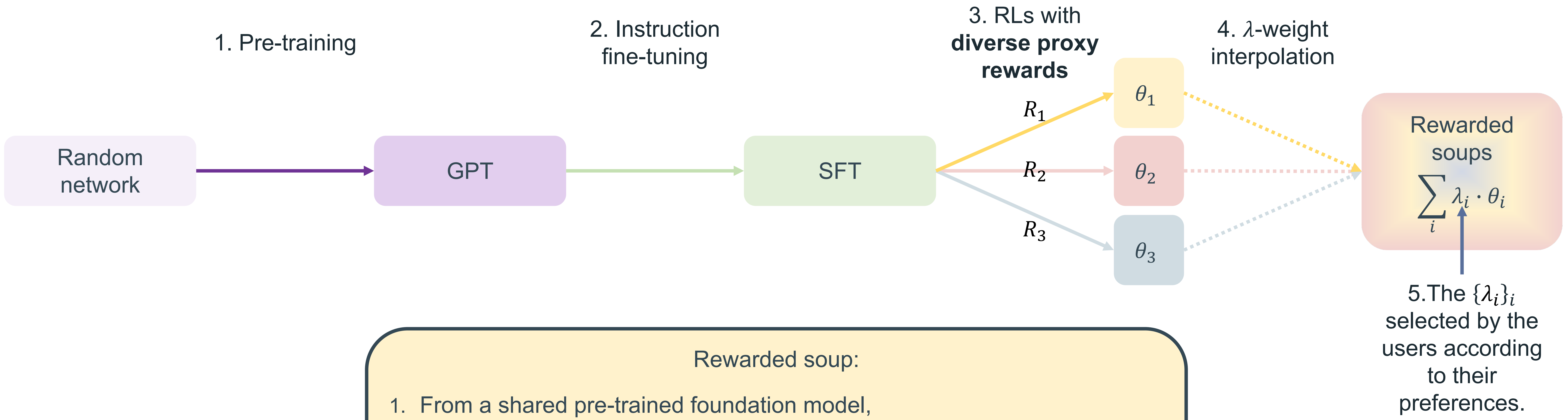
- **Safety**: helpfulness vs harmlessness?
- **Summarization**: complete or factual ?

Diversity of people and applications ⇒ which one should we optimize for?

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

Move from a single-policy towards a **multi-policy paradigm** to embrace diversity.

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

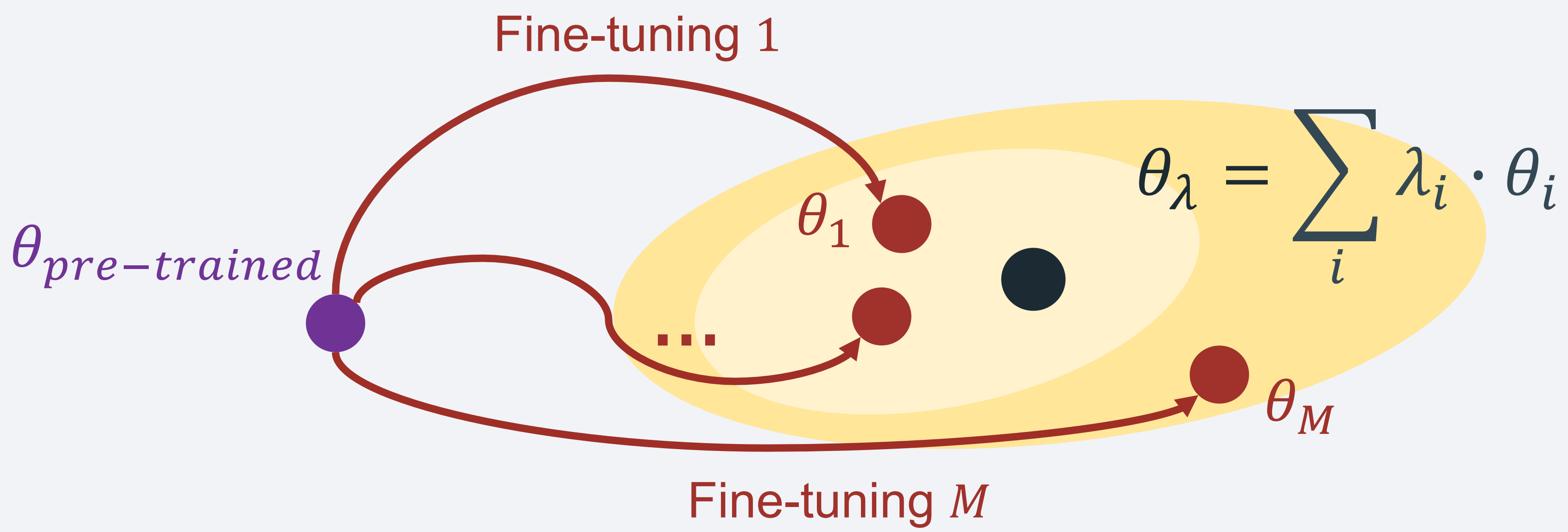


Rewarded soup:

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

Weight interpolation relies on linear mode connectivity

When fine-tuned from a shared **pre-trained** model, weights remain **linearly connected** and thus can be interpolated despite the non-linearities in the architecture.



[Wortsman2022] Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. ICML.
[Rame2023] DiWA: diverse weight averaging for out-of-distribution generalization. NeurIPS.

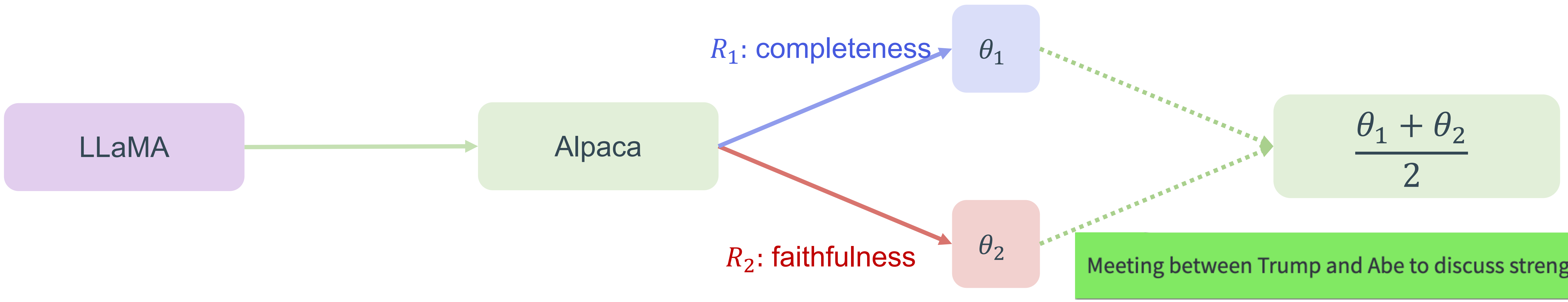
Summarization with diverse reward models

Trump-Abe meeting

Text to summarize:

A meeting in New York on Thursday between U.S. President-elect Donald Trump and Japanese Prime Minister Shinzo Abe has ended, a Trump transition team official said. The hastily arranged meeting was an attempt to smooth relations following Trump's campaign rhetoric that cast doubt on long-standing U.S. alliances.

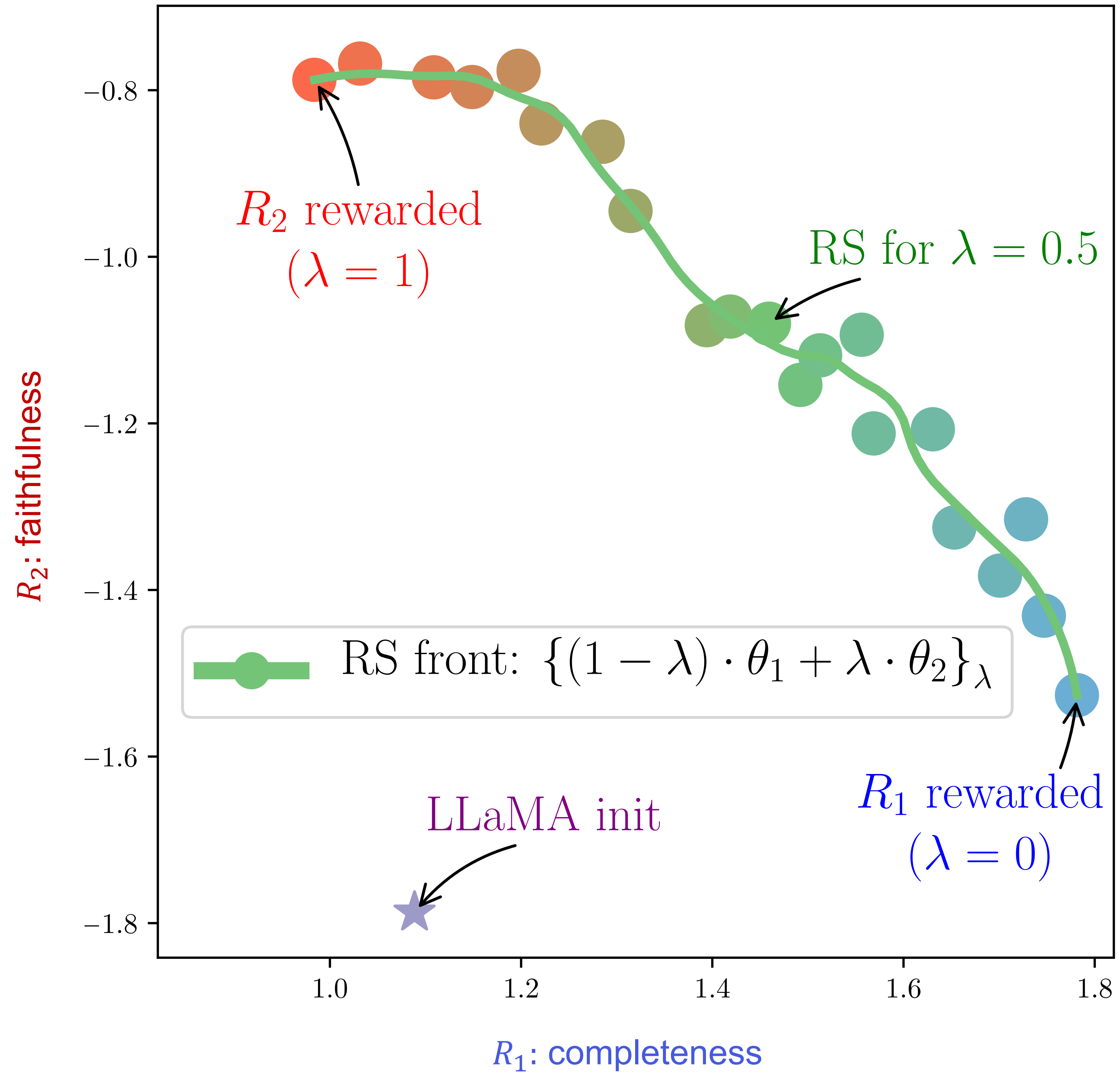
Meeting between Trump and Abe in New York to discuss strengthening US-Japan alliance following Trump's campaign rhetoric that cast doubt on long-standing



Meeting between Trump and Abe to discuss strengthening ties between the US and Japan.

Meeting between Trump and Japan PM to smooth relations.

Summarization with diverse reward models



Trump-Abe meeting

Text to summarize:

A meeting in New York on Thursday between U.S. President-elect Donald Trump and Japanese Prime Minister Shinzo Abe has ended, a Trump transition team official said. The hastily arranged meeting was an attempt to smooth relations following Trump's campaign rhetoric that cast doubt on long-standing U.S. alliances.

Generated summaries:

λ=0.0

Meeting between Trump and Abe in New York to discuss strengthening US-Japan alliance following Trump's campaign rhetoric that cast doubt on long-standing

λ=0.5

Meeting between Trump and Abe to discuss strengthening ties between the US and Japan.

λ=1.0

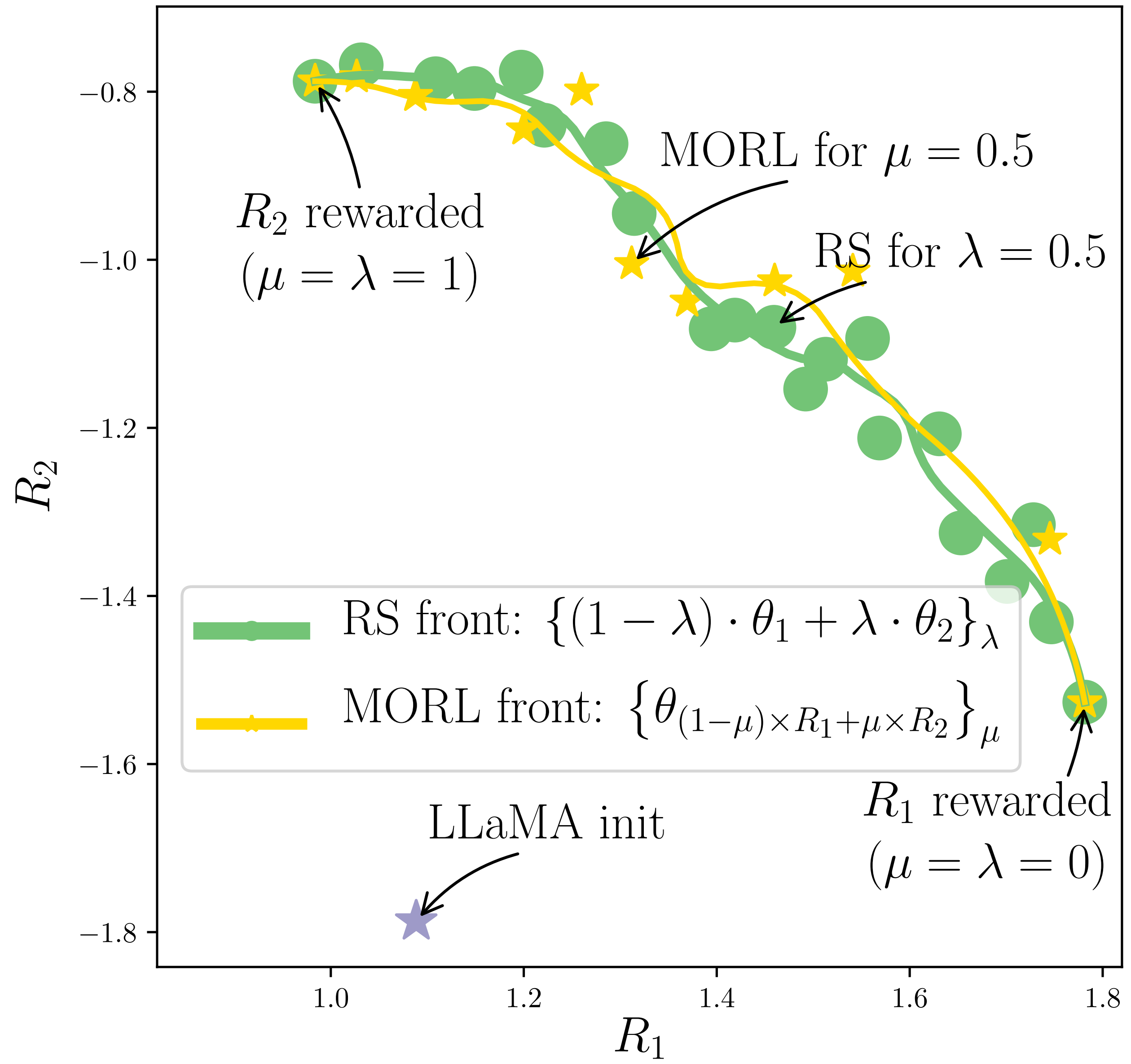
Meeting between Trump and Japan PM to smooth relations.

Pareto-optimal alignment across rewards

Rewarded soups

Interpolate the weights a posteriori:

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



Multi-objective: MORL

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.

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

We apply rewarded soup in multiple standard learning tasks:

1. Text

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

2. Multimodal: text and image

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

3. Continuous control

- Locomotion.

Benefits from rewarded soups

1. Efficiency

- 1 single fine-tuning per reward.
- Parralelization.
- No inference overhead.

2. Transparence

- Support decision-making.
- Facilitate regulation by an (external) non-technical committee.

3. Updatable

- Easily update the λ.
- Easily add new reward.
- Iterative development process.

4. Fairness

- Value pluralism.
- Under-represented groups.
- Less ideological hegemony.
- Federated learning setups?.

Conclusion

- Human-aligned AI as a multi-objective problem.
- Weight interpolation as a practical pareto-optimal solution.
- Code: https://github.com/alexrame/rewardedsoups