Snorkel

Programmatically scaling human preferences and alignment in GenAI

January 2024

Hoang Tran
Applied Research Scientist, Snorkel AI

Snorkel

Key Takeaways

To optimize your model responses:

1. **Further Alignments:** After fine-tuning with high-quality data, consider additional alignments using techniques such as RLHF or DPO to further fit your preferences

2. **Human Preferences Data:** Both RLHF and DPO require human preferences data, which can be resource-intensive in enterprise settings

3. **Scalable Solution**: Streamline the human preferences process by programmatically scaling it with weak supervision, reducing the time and resources required

Snorkel

Agenda

1. Aligning LLMs with human preferences: RLHF and DPO

2. Efficiently scale SMEs preferences: The programmatic approach

3. Results

Part 1: Aligning LLMs with human preferences: RLHF and DPO

Snorkel

ChatGPT Moment

November, 2022 - ChatGPT has taken the world by storm thanks to its advanced conversational capabilities

2 months after, ChatGPT reaches 100 Million Users

What initially distinguishes ChatGPT and makes it a “likeable” choice among users?

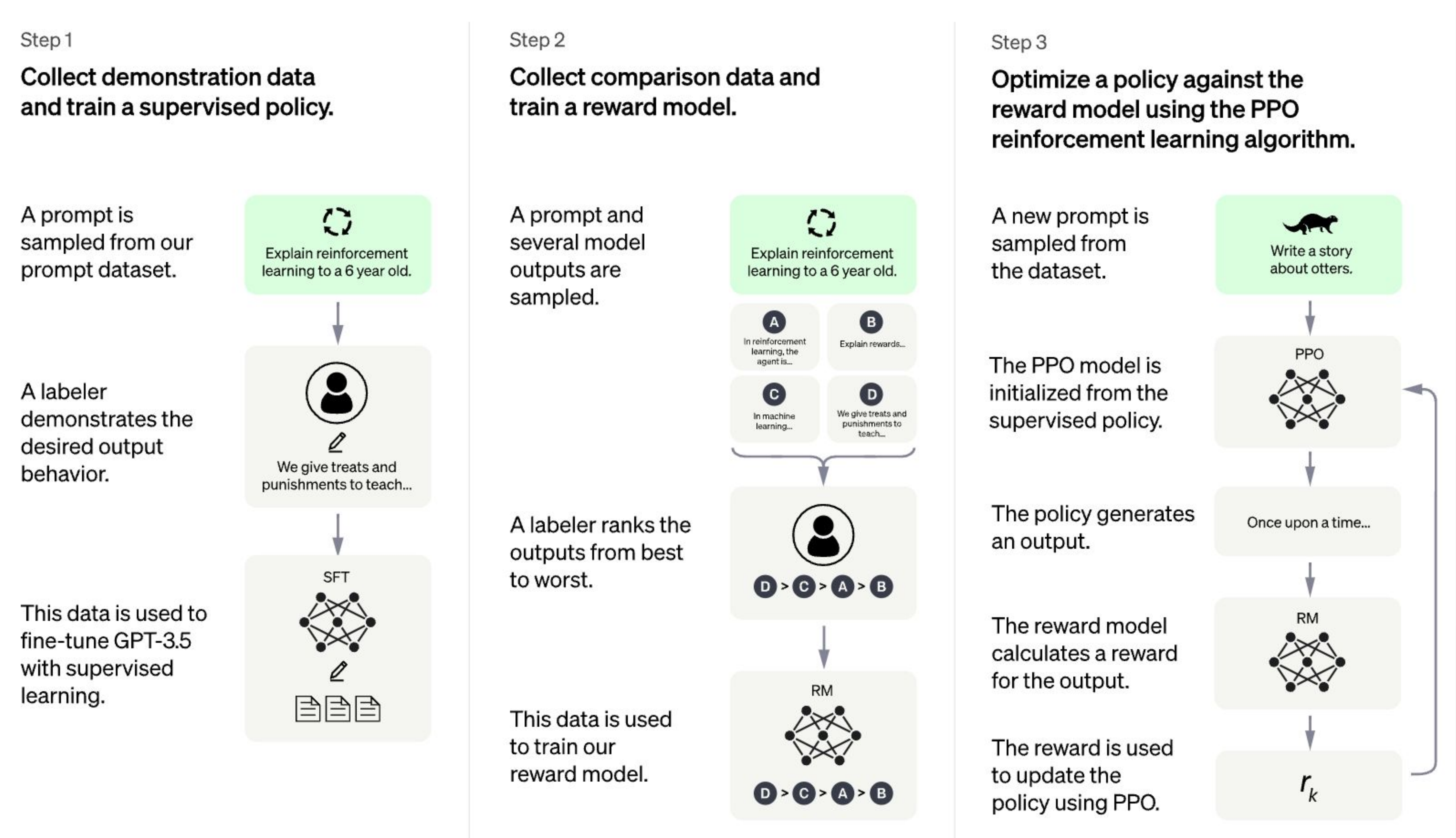
ChatGPT Recipe

Supervised Fine-tuning (Instruction Tuning)

Reward Modeling

Reinforcement Learning from Human Feedback (RLHF)

Pretraining

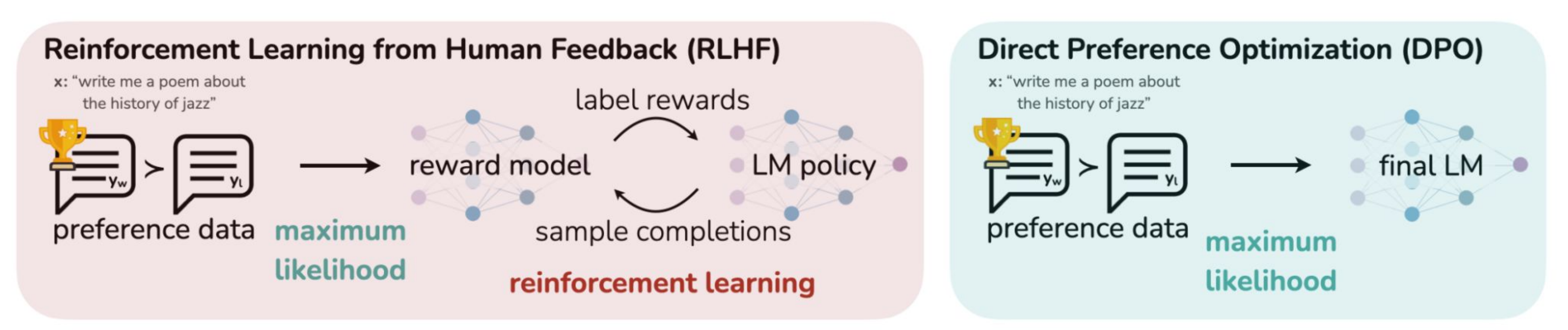


Source: OpenAI, 2022

Alignment steps are crucial to develop LLM that follows your enterprise customized preferences

Snorkel

Alternative to RLHF: Direct Preference Optimization



Similarity: DPO and RLHF both utilize a preference dataset

Differences:

- DPO skips the creation of a reward model and reinforcement learning iterations
- DPO update increases the relative log probability of chosen to rejected responses
  → Aiming to generate responses **closer to chosen texts** and **further from rejected texts**

Note: This talk is NOT about DPO vs RLHF or RL, but about how to scalably create the preference dataset.

Source: https://arxiv.org/pdf/2305.18290.pdf Rafailov, Sharma, Mitchell, et. al., 2023

DPO vs RLHF: a limited comparison

Computation

DPO is computationally lighter, eliminating the need for sampling from the LLM during training compared to RLHF

Exploration

RLHF supports more exploration as it is only constrained by reward scores (with KL penalty), while DPO directly optimizes against preferred text

Output Models

RLHF with GPT, Claude, Llama, Gemini

DPO has shown promising theoretical results, contributing to Zephyr, Tulu v2*

* Note: Zephyr, Tulu DPO recipe use responses from other LLMs

More experiments are needed to comprehensively compare different approaches

BUT the common theme: **Both approaches require data that reflects preferences**

Snorkel

Preference data caveats

Resources

Collecting annotations for building preference datasets is **resource-intensive and time-consuming**

Iterative Process

As LLMs improves after updates, you need to **continuously collect more updated preference data** and update your model (weekly - Llama 2)

Enterprise Customization

Current alignment axis are simple: helpful, honest, harmless, safety.

The challenge lies in scaling alignments to **advanced enterprise customization** to reflect internal policies and preferences

You need scalable way to create your customized preference data

Part 2: Efficiently scale SMEs preferences: The programmatic approach

The end-goal

Datasets that reflect human/subject-matter-experts (SMEs) preferences:

1. **Pairwise/Ranking preferences**
   Between these 5 responses, which response do you like best? Rank the responses.
   Requirements: Multiple responses for the same prompt

2. **High/Low-quality classification**
   Classifying/labeling each response are high/low-quality (or of varying ratings)
   Multiple responses for the same prompt is **optional**

Involving SMEs in the Data Development Process

SMEs need to be closely in the loop to align
the data development process with **business needs**

Manual Annotation

Engage SMEs in providing a **small subset** of manual annotations for validation

Knowledge Base

Connect with **knowledge base** to enhance contextual understanding

Programmatic Labelling

Instead of manually labelling one-by-one, create labels with **functions:** desired text patterns, formats, prompts, external metrics, models, etc.

Snorkel will **combine and denoise various signals** to provide quality dataset

Scalable Data Development with SMEs

Instead of manually labelling one-by-one, allow SMEs to **create labels with functions** that express their preferences **at scale**

Examples can include:

1. Text patterns

2. Prompting with LLMs

3. External specialized models and metrics

These labelling functions can be noisy and may conflict.
Snorkel Flow will combine and denoise the various signals to generate a quality labelled training set

Scalable Data Development with SMEs:
1. Text patterns

Some examples: (we support more fine-grained customization)

- Prefer responses with:
  - **Format:** Prefer list-like responses, with follow-up questions
  - **Marketing purposes:** Prefer responses that mention the company name at least X times
  - **Workflow adherence:** The ideal conversations need to follow 3 steps

- Downgrade responses with:
  - **Safety:** Remove responses sensitive words
  - **Format:** Disprefer responses that immediately answer a question with a question
  - **Pattern:** Disprefer responses with high adjective ratio (longer, more descriptive, rather than direct response)

Scalable Data Development with SMEs:
2. Prompting with LLMs

In addition to label with text patterns, SME can prompt LLMs to support labelling

Common themes:

- External ratings from performant LLMs
- Align to more arbitrary, generic characteristics like workflow adherence, safety, helpfulness
- Direct QnA on your data to support decision making

Snorkel

Scalable Data Development with SMEs:
3. External specialized models and metrics

Utilize existing specialized model and metrics:

- **External metrics:** low perplexity, low toxicity scores, high sentiment scores
- **External models:** Utilize signals from existing models built from external datasets: e.g., Open Assistant chat dataset, FinGPT model
- **External features:** Create additional features like key topics, supportive text from databases, past ratings

Scalable Data Development with SMEs

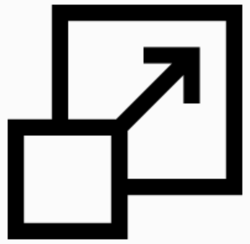
The above functions are only examples, and they can be noisy and conflicting. **Snorkel Flow will combine and denoise the various signals to generate a quality labelled training set**

We supports **more customization** to build your **custom models on your data**

Snorkel

The collaboration of SMEs and Snorkel techniques is crucial to scalably achieve data reflective of enterprise preferences

Benefits of Snorkel data development



Scalable

Simultaneous label multiple data points, increase efficiency

45x faster in a case study with a Fortune 50 Bank*

→ **Support:** resources constraints & allow more customization

Tractable

Labels developed with labelling functions are **traceable to their origin**, enhancing **auditability** and making **error correction more tractable**

→ **Support:** enterprise-level auditability and iteratively improving data

Transferable

You can **transfer existing signals** into new development phases
→ Developing with updated data or different alignment axes is faster

→ **Support:** bootstrap iterative data process & save resources

* Source: https://snorkel.ai/case-studies/

Part 3: Results

Build reward model with unlabelled data

In July 2023, we **programmatically labelled high/low quality responses** from Dolly & Open Assistant open-sourced data.

Two developers accomplished this programmatically in 1 day, avoiding the need for weeks or months of manual annotation

Then using the labels, we **develop a quality-scoring/reward model*** to classify and scores if a given prompt-response is high or low quality

For more info, visit: https://snorkel.ai/how-we-built-better-genai-with-programmatic-data-development/

Snorkel

Build reward model with unlabelled data

Example functions to label high vs low quality for a generic chat LLM:

- **Text pattern:**
  - Prefer **list-like** responses, saying thank you/positive responses
  - Downgrade when LLMs answer users with questions

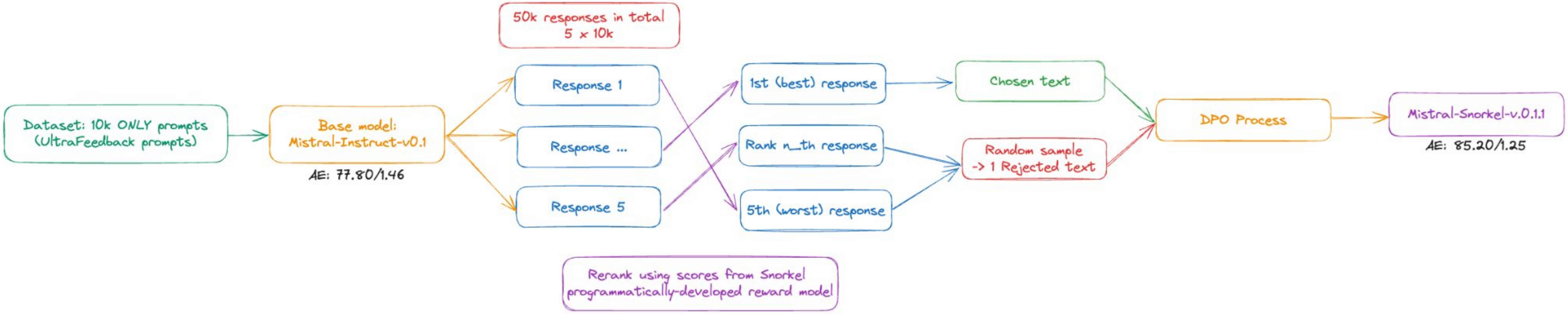
- **External metrics:**
  - Perplexity: Prefer **low perplexity** for creative text generation tasks
  - Embeddings: **High cosine similarity** between questions & responses

- **Task-dependent patterns:**
  - We also **programmatically developed a model to classify task types** (QnA, chat, summarization, etc.) and use these tags to control for **task-dependent quality** E.g., length for summarization tasks, conversation patterns for chat tasks

Source: https://snorkel.ai/how-we-built-better-genai-with-programmatic-data-development/

Snorkel

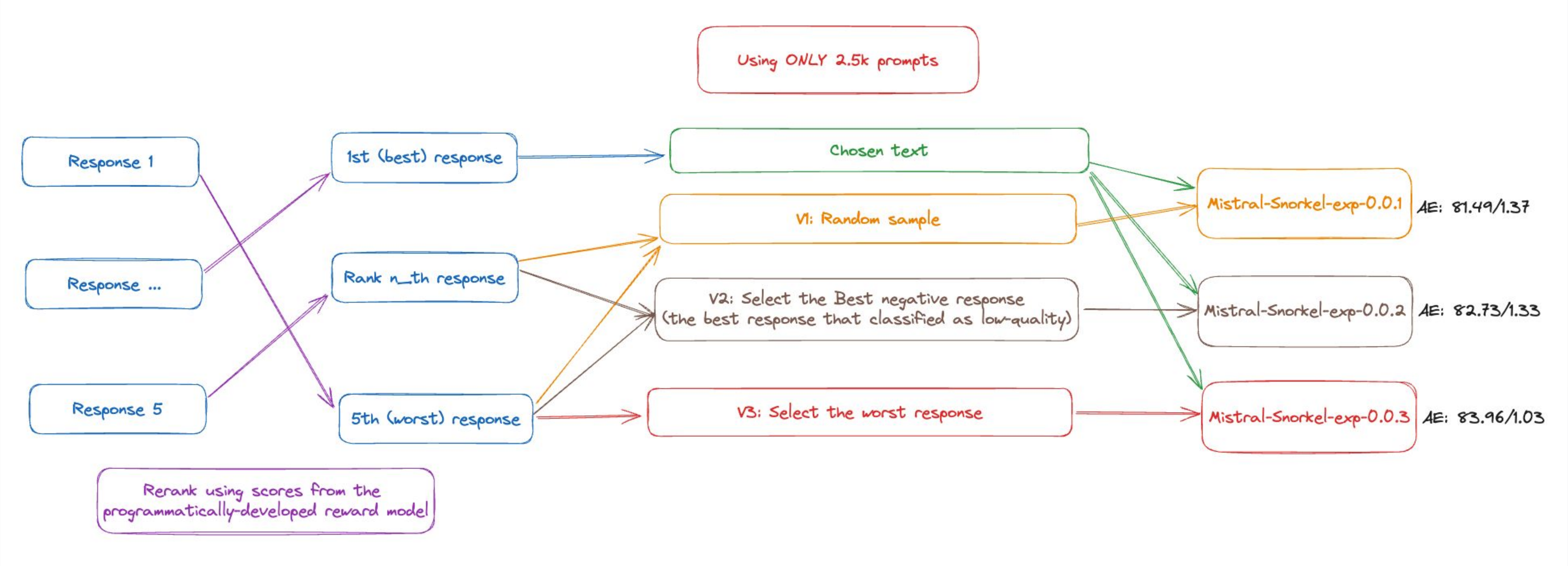
DPO on self-generated responses with preferences provided by Snorkel reward model



Key results:

- **7.4 points (9.5%) increase** on Alpaca-Eval (77.80 → 85.20)
- Developed **under 1 day** & NOT using outputs from other LLMs

The data selection effect



Key results:

- Hard negative sampling (v3 - 83.96) performs better than random sampling (v1 - 81.49)
- Access to relative scores is useful for optimizing performance and customization

Snorkel

Key results

With a **10k prompt-only dataset** and **no learning on responses from other LLMs**:

- Achieved a **7.4-point (9.5%) increase** on Alpaca-Eval in **under 1 day** (77.80 → 85.20)

- Competitive against alternative:
  - DPO with preference from Snorkel model: 85.20
  - DPO with preference from Open Assistant model: 83.31

- LLMs can improve through DPO on its self-generated responses with **preference scores from external models customized** to reflect your enterprise preferences

Note: Despite the competitive results, the ultimate goal of the above experiments is **NOT to compete on the public benchmark** but to **mimic our data-centric approach**, replicating perspectives and results observed in **our engagements with leading F500 enterprises**.
For more details, visit https://snorkel.ai/first-snorkel-foundry-cohort-achieves-gains-of-up-to-54-points/

Snorkel

Key results

When DPO, hard negative sampling (best vs worst) **performs better** than random sampling
→ Need to collect more comprehensive rankings or utilize reward model scores

Additional benefits: The reward model can help **select high-quality data** for supervised fine-tuning

High-quality supervised fine-tuning data are important (LIMA by Meta, 2023)

Utilize the same reward model, we select a subset of high-quality data to train *RedPajama-7B-Chat-Curated*, which **outperforms 3.5 to 10 points** against when trained on all data

Source: https://snorkel.ai/how-we-built-better-genai-with-programmatic-data-development/

In enterprise settings

At a Fortune 500 telecommunications company:

- **Use Case:** Chatbot/Co-pilot

- Involving SMEs in the loop to **programmatically** identify **high-quality** customer-support responses that **adheres to internal workflows** (data-quality and customization)

- **Result:** Achieved a **+17 point boost** on F1 score in predicting high-quality and workflow-adhered responses, outperformed using ChatGPT alone in predicting high quality

* Source: https://snorkel.ai/first-snorkel-foundry-cohort-achieves-gains-of-up-to-54-points/

Key Takeaways

To optimize your model responses:

1. **Further Alignments:** After supervised fine-tuning, perform additional alignments for point boosts and **enhanced enterprises policies adherence**

2. **Human Preferences:** Developing LLMs to fit enterprise policies involves a **costly and time-consuming** iterative data development process.

3. **Snorkel technology and data-centric workflow**: Develop customized preference data with **scalability, tractability, and transferability.**
   Build data with **SMEs' in-the-loop** to train LLMs more closely with enterprise preferences

Snorkel

Thank you!

Hoang Tran
hoang.tran@snorkel.ai/ LI: Viet Hoang Tran Duong