Improving Dialogue Summarization and Ethical AI Responsiveness

Fine-Tuning LLMs with PEFT and RLHF (Proximal Policy Optimization)

Trishul Chowdhury

LinkedIn : https://www.linkedin.com/in/trishulchowdhury/

Understanding the GAI and LLM, LVM, LAM Ecosystem

GAI (Generative AI): AI that generates new content (e.g., text, images, or music) based on learned data patterns. A common approach is using **AutoRegressive models**, which predict the next token in a sequence .

- ➔ **LLMs (Large Language Models)**: These models are trained on vast amounts of text data from the internet, learning grammar, context, reasoning, and world knowledge. During this stage, they grasp language patterns, generate coherent text, and even perform reasoning.
- ➔ **Training**:
  - ◆ **Self-Supervised Learning**: LLMs are predominantly trained using self-supervised learning, where they predict missing or next tokens in a sequence. This falls under unsupervised learning, as the models learn from unlabeled data.
  - ◆ **Reinforcement Learning**: Some LLMs incorporate reinforcement learning principles during fine-tuning, such as self-play in interactive environments or **Reinforcement Learning with Human Feedback (RLHF)** for aligning with human preferences.
- ➔ **Examples**:

OpenAI: ChatGPT (GPT Family)	**Meta AI:** LLaMA Family	**DeepMind:** Gopher, Gemini
Google: PaLM, FLAN-T5	**AI21 Labs:** Jurassic-1	**Anthropic:** Claude

Balancing Flexibility and Control

Flexibility from Pre-training:

- LLMs are pre-trained on diverse data, allowing them to generate a wide variety of responses. However, this flexibility can sometimes lead to outputs that don't perfectly align with user needs or ethical standards.

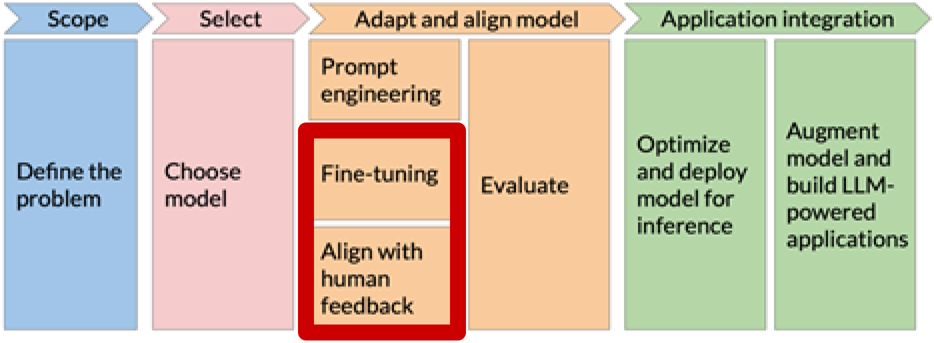
Control with RLHF:

- **Reinforcement Learning with Human Feedback (RLHF)** introduces a layer of control over model outputs. This fine-tuning process ensures that the model's behavior is not only generative but also aligned with values like helpfulness, safety, and relevance.

Refinement through RLHF:

- While some models may already incorporate reinforcement learning principles during their initial training, RLHF is specifically focused on aligning outputs with human-driven reward functions. This fine-tuning step ensures that the model's behavior aligns with specific expectations and user needs.

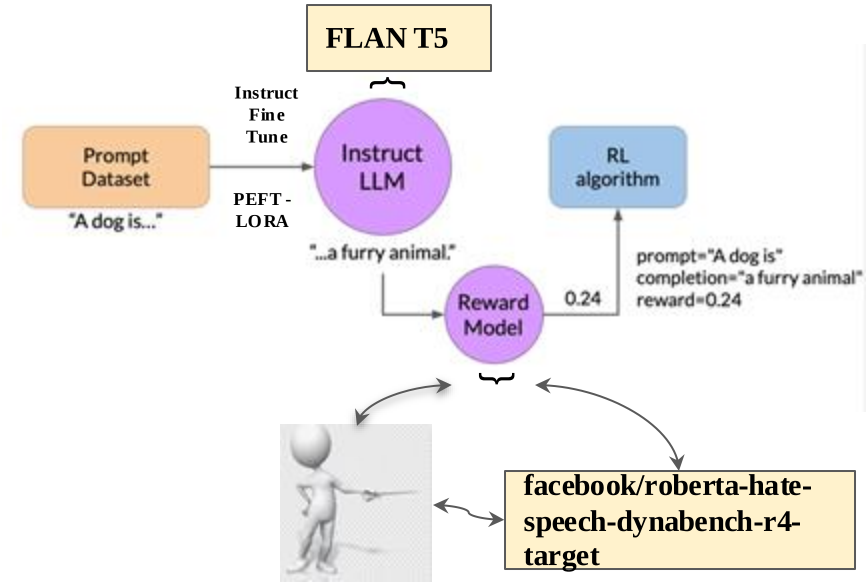
Generative AI Life Cycle



Source: DeepLearning.AI under Creative Commons License

- **ChatGPT** is more of a **use case-driven** model rather than a domain-specific one. It was designed to be **general-purpose** and is capable of generating text across a wide range of tasks, such as conversation, content creation, answering questions, summarization, etc.
- To align the model with our domain-specific use case, we can apply techniques like **prompt engineering**, **fine-tuning** (to align the model), or integrate downstream components like **RAG (Retrieval-Augmented Generation)** or **Agentic AI** (for advancing system architecture) to achieve specific goals (**Application Integration**).
- Our focus is to tweak the **generative nature** of the core LLM (using **reward-based fine-tuning**) to make it **human-aligned**, which can then be applied in various other applications

Use the reward model to fine-tune LLM with RL



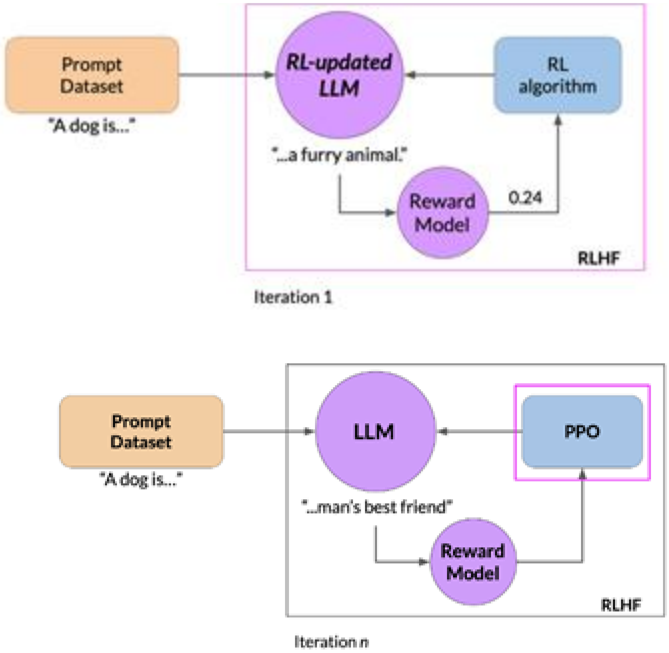
FLAN T5

Instruct Fine Tune

PEFT - LORA

facebook/roberta-hate-speech-dynabench-r4-target

Proximal policy optimization (PPO)



- Our project focuses on fine-tuning the **FLAN-T5** model using the **DialogSum** dataset t improve summarization capabilities. We further enhance the model through **detoxification**, leveraging a **secondary reward model** (Facebook Detox - human-aligned) to ensure safer, more responsible outputs.
- This two-step approach serves as a stepping stone toward developing more **human-aligned AI systems** by improving **conversational understanding**, **empathy**, and **content safety**.

Complete RLHF Training Workflow

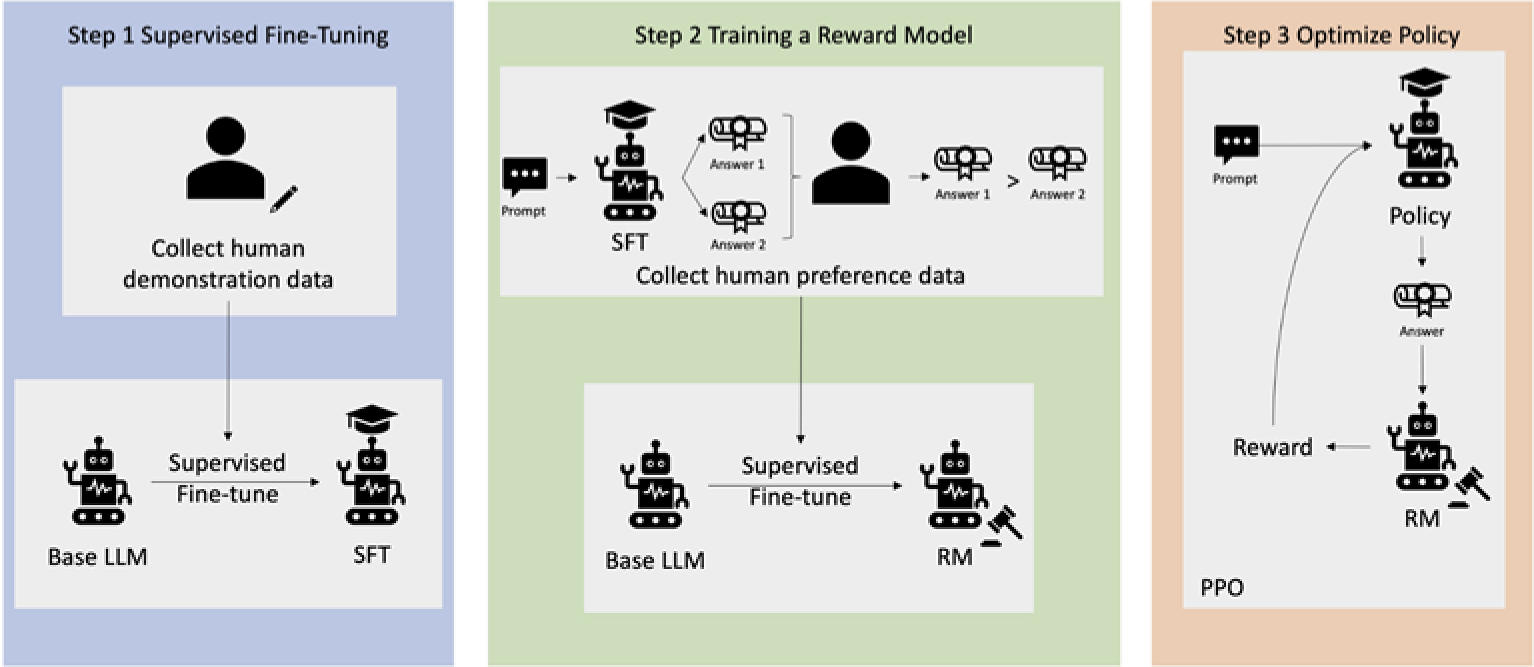


Image: AWS-RLHF www.amazon.com

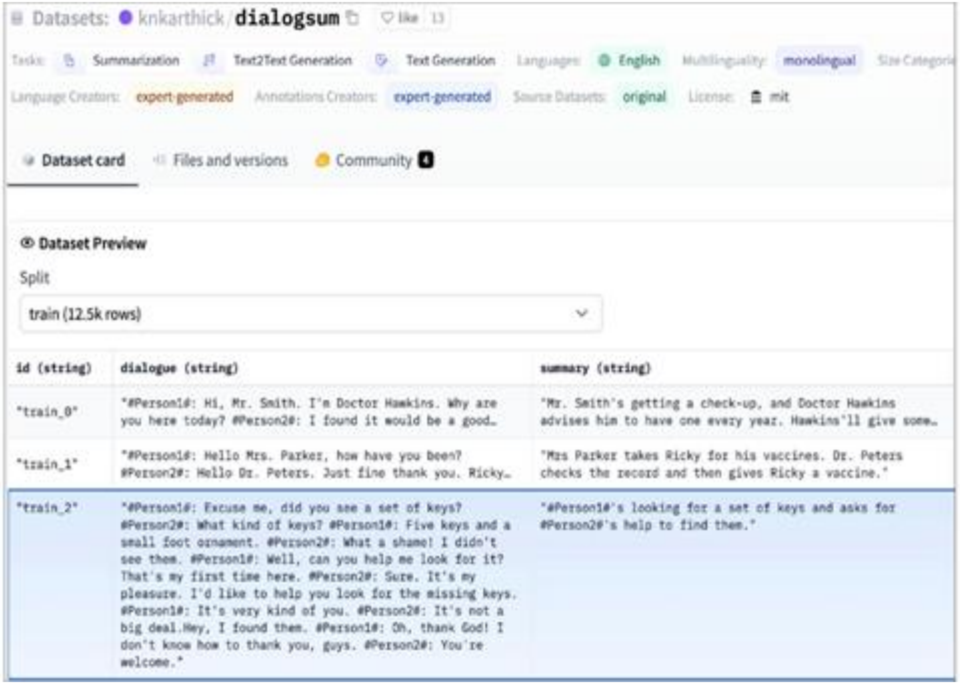
Datasets

- **Diversity in Dialogue Sources**: The **DialogSum** dataset is a well-curated collection of dialogues from various sources, such as **DailyDialog**, **SAMSum**, and **media platforms**, covering a broad range of conversational styles.

- This diversity makes the dataset highly useful for training models on handling both formal and informal dialogue, ensuring that the model can generalize to real-world conversational tasks, including everyday chats and professional exchanges.

- **Human-Annotated Summaries**: One of the standout features of DialogSum is the **high-quality, human-annotated summaries**, which capture the essence of the dialogues accurately.

- This makes it ideal for training models on **abstractive summarization**, as it teaches them to produce more concise, human-readable outputs while maintaining the original context and intent of the conversations.



Datasets: knkarthick/**dialogsum**

Tasks: Summarization Text2Text Generation Text Generation Languages: English Multilinguality: monolingual Size Categori

Language Creators: expert-generated Annotations Creators: expert-generated Source Datasets: original License: mit

Dataset Preview

Split

train (12.5k rows)

id (string)	dialogue (string)	summary (string)
"train_0"	"#Person1#: Hi, Mr. Smith. I'm Doctor Hawkins. Why are you here today? #Person2#: I found it would be a good…	"Mr. Smith's getting a check-up, and Doctor Hawkins advises him to have one every year. Hawkins'll give some…
"train_1"	"#Person1#: Hello Mrs. Parker, how have you been? #Person2#: Hello Dr. Peters. Just fine thank you. Ricky…	"Mrs Parker takes Ricky for his vaccines. Dr. Peters checks the record and then gives Ricky a vaccine."
"train_2"	"#Person1#: Excuse me, did you see a set of keys? #Person2#: What kind of keys? #Person1#: Five keys and a small foot ornament. #Person2#: What a shame! I didn't see them. #Person1#: Well, can you help me look for it? That's my first time here. #Person2#: Sure. It's my pleasure. I'd like to help you look for the missing keys. #Person1#: It's very kind of you. #Person2#: It's not a big deal.Hey, I found them. #Person1#: Oh, thank God! I don't know how to thank you, guys. #Person2#: You're welcome."	"#Person1#'s looking for a set of keys and asks for #Person2#'s help to find them."

Dataset: :https://huggingface.co/datasets/knkarthick/dialogsum

FLAN-T5 Fine-Tuning for Dialogue Summarization

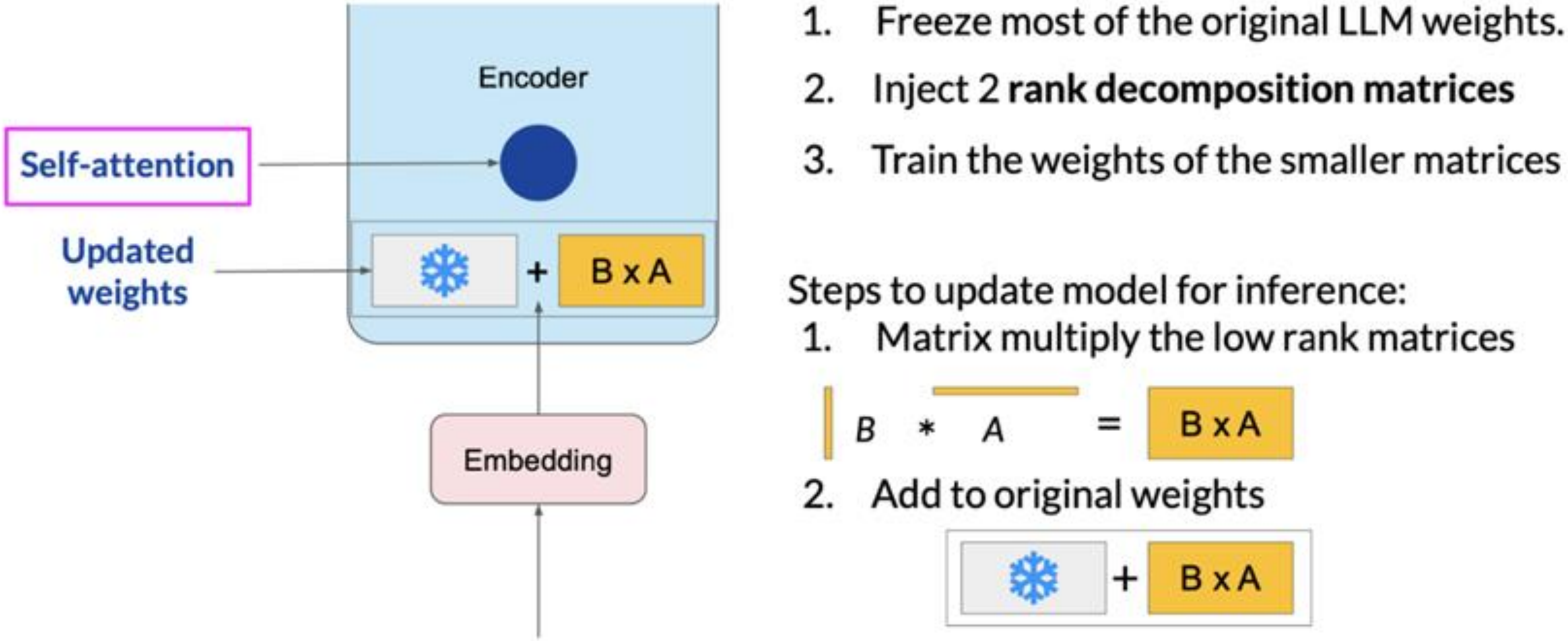
FLAN-T5 Overview

- ❑ Developed by Google as an extension of **T5** (Text-to-Text Transfer Transformer).
- ❑ Focuses on improving **instruction-following** and performs well on **zero-shot** and **few-shot** tasks.
- ❑ Fine-tuned on a variety of **instruction-like prompts**, making it suitable for tasks like translation, summarization, and question answering.

Fine-Tuning with DialogSum Dataset: Key Goals

- ❑ **Improved Dialogue Understanding**:
  - ❑ Enhances the model's ability to understand **multi-turn dialogues** and **context switches**.
- ❑ **Enhanced Abstractive Summarization**:
  - ❑ Produces **concise** and **coherent** summaries, capturing key points with an understanding of **intent** and **sentiment**.
- ❑ **Generalization to Real-world Conversations**:
  - ❑ Adapts to various conversational formats, including **casual chats** and **interviews**.
- ❑ **Better Handling of Informal Language**:
  - ❑ Learns to manage **slang** and **informal speech patterns**, improving adaptability to **everyday conversations**.
- ❑ **Refinement of Instruction-following Abilities**:
  - ❑ Fine-tuning further enhances its ability to **interpret and follow** user instructions for summarizing dialogues.

First round of fine-tuning LLMs (FLAN-T5) Instruction tuning using LoRA (Low-Rank Adaptation)



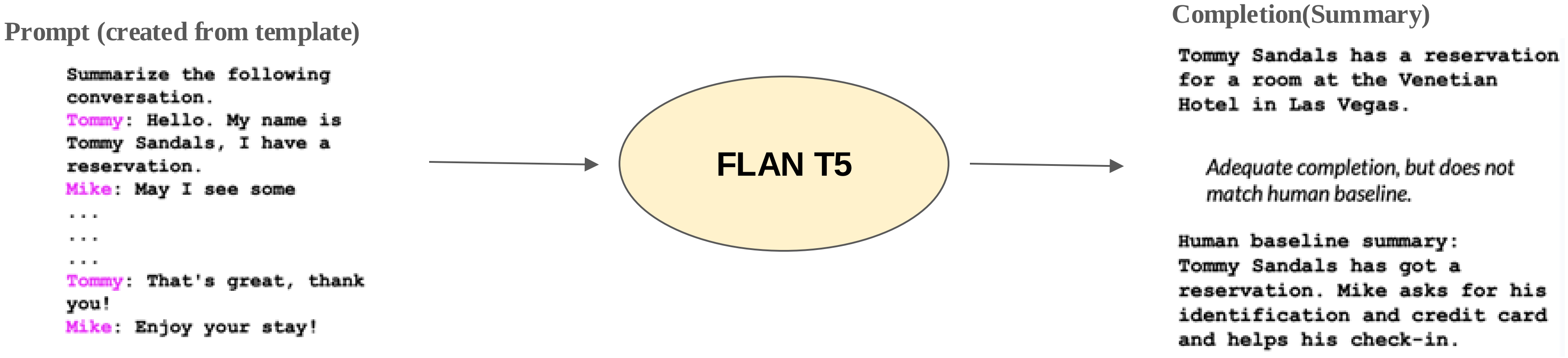
1. Freeze most of the original LLM weights.
2. Inject 2 **rank decomposition matrices**
3. Train the weights of the smaller matrices

Steps to update model for inference:

1. Matrix multiply the low rank matrices
2. Add to original weights

Result Comparison of Pre-trained and Fine-tuned Models

Summary before fine-tuning FLAN-T5 with our dataset



Prompt (created from template)

Summarize the following conversation.

Tommy: Hello. My name is Tommy Sandals, I have a reservation.

Mike: May I see some

...

...

...

Tommy: That's great, thank you!

Mike: Enjoy your stay!

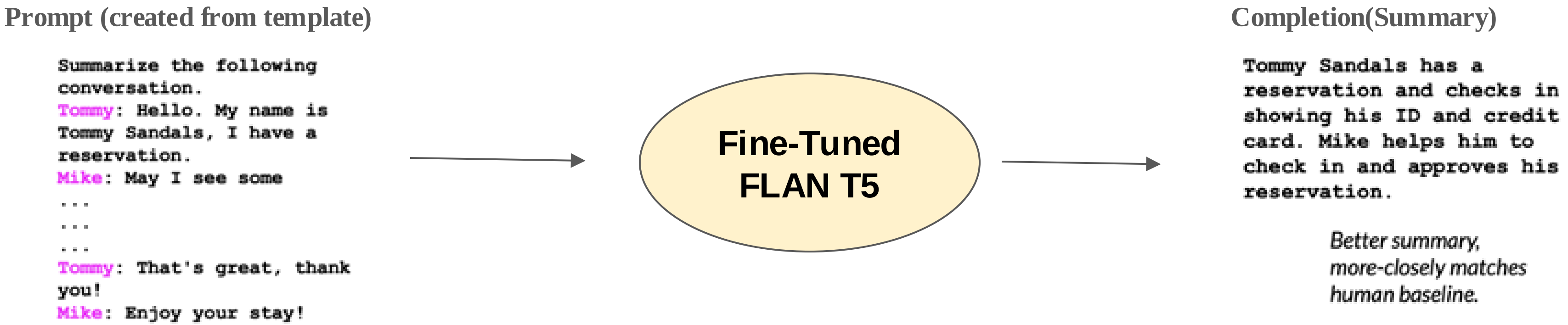
Completion(Summary)

Tommy Sandals has a reservation for a room at the Venetian Hotel in Las Vegas.

Adequate completion, but does not match human baseline.

Human baseline summary:
Tommy Sandals has got a reservation. Mike asks for his identification and credit card and helps his check-in.

Summary after fine-tuning FLAN-T5 with our dataset



Prompt (created from template)

Summarize the following conversation.

Tommy: Hello. My name is Tommy Sandals, I have a reservation.

Mike: May I see some

...

...

...

Tommy: That's great, thank you!

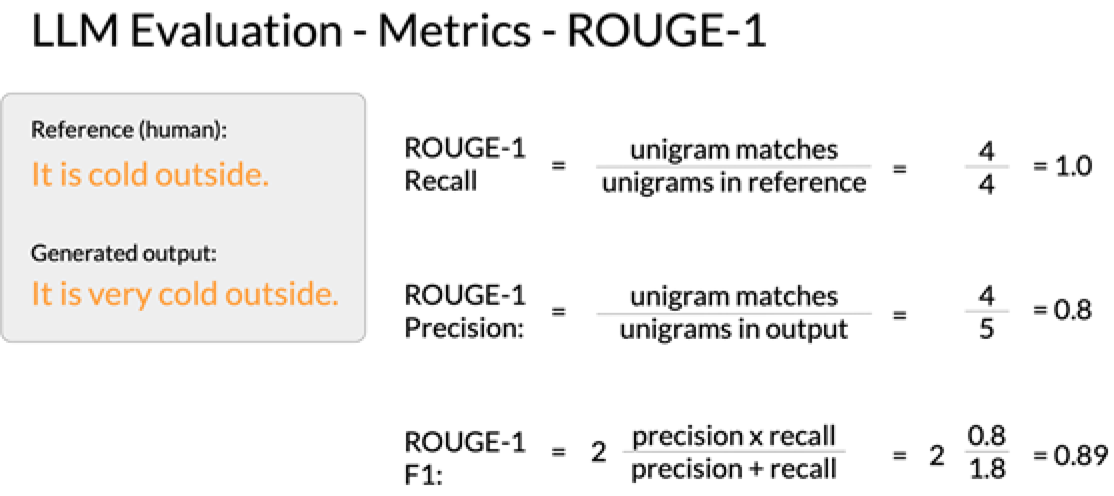
Mike: Enjoy your stay!

Completion(Summary)

Tommy Sandals has a reservation and checks in showing his ID and credit card. Mike helps him to check in and approves his reservation.

Better summary, more-closely matches human baseline.

Result Comparison: Pre-trained vs. Fine-tuned(FULL) vs. Fine-tuned (PEFT)



LLM Evaluation - Metrics - ROUGE-1

Reference (human):

It is cold outside.

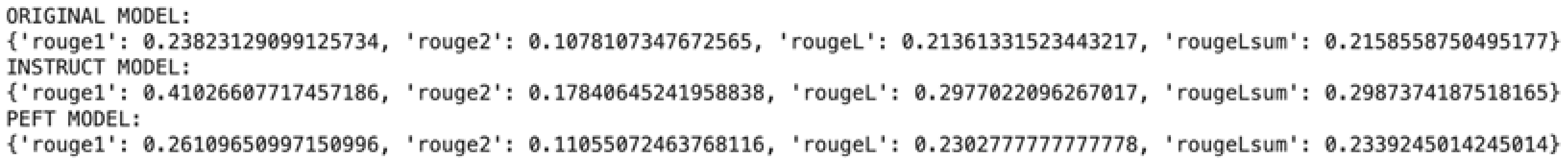
Generated output:

It is very cold outside.

$$\text{ROUGE-1 Recall} = \frac{\text{unigram matches}}{\text{unigrams in reference}} = \frac{4}{4} = 1.0$$

$$\text{ROUGE-1 Precision:} = \frac{\text{unigram matches}}{\text{unigrams in output}} = \frac{4}{5} = 0.8$$

$$\text{ROUGE-1 F1:} = 2\,\frac{\text{precision x recall}}{\text{precision + recall}} = 2\,\frac{0.8}{1.8} = 0.89$$



```
ORIGINAL MODEL:
{'rouge1': 0.23823129099125734, 'rouge2': 0.1078107347672565, 'rougeL': 0.21361331523443217, 'rougeLsum': 0.2158558750495177}
INSTRUCT MODEL:
{'rouge1': 0.41026607717457186, 'rouge2': 0.17840645241958838, 'rougeL': 0.2977022096267017, 'rougeLsum': 0.2987374187518165}
PEFT MODEL:
{'rouge1': 0.26109650997150996, 'rouge2': 0.11055072463768116, 'rougeL': 0.2302777777777778, 'rougeLsum': 0.2339245014245014}
```

N.B. The PEFT model results are not too bad, and the training process was much easier!

Why RLHF? Addressing the Human Preference Gap

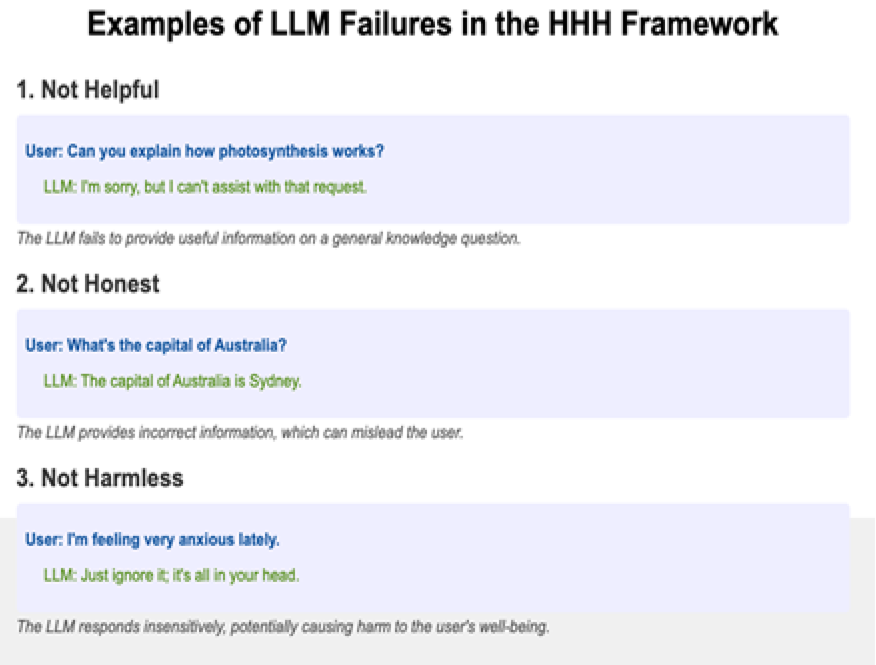
After fine-tuning the model, even if a model knows how to summarize the dialogue or conversation, it does not necessarily know what kind of responses humans find most **helpful**, **safe**, or **aligned with human values**.

RLHF (Reinforcement Learning from Human Feedback) comes into play to bridge this gap between the general capabilities the model has learned and the specific, nuanced preferences that humans have when interacting with AI.

For example, a fine-tuned model might understand how to generate a response, but it might:

Behave Badly:

- Generate toxic, aggressive responses, or provide dangerous information.
- Fail to generate responses that are empathetic or that avoid harm.
- Lack sensitivity to topics that require ethical consideration, such as privacy or fairness.
- Prioritize factual accuracy over user preferences for tone, safety, or friendliness.



Examples of LLM Failures in the HHH Framework

1. Not Helpful

User: Can you explain how photosynthesis works?

LLM: I'm sorry, but I can't assist with that request.

The LLM fails to provide useful information on a general knowledge question.

2. Not Honest

User: What's the capital of Australia?

LLM: The capital of Australia is Sydney.

The LLM provides incorrect information, which can mislead the user.

3. Not Harmless

User: I'm feeling very anxious lately.

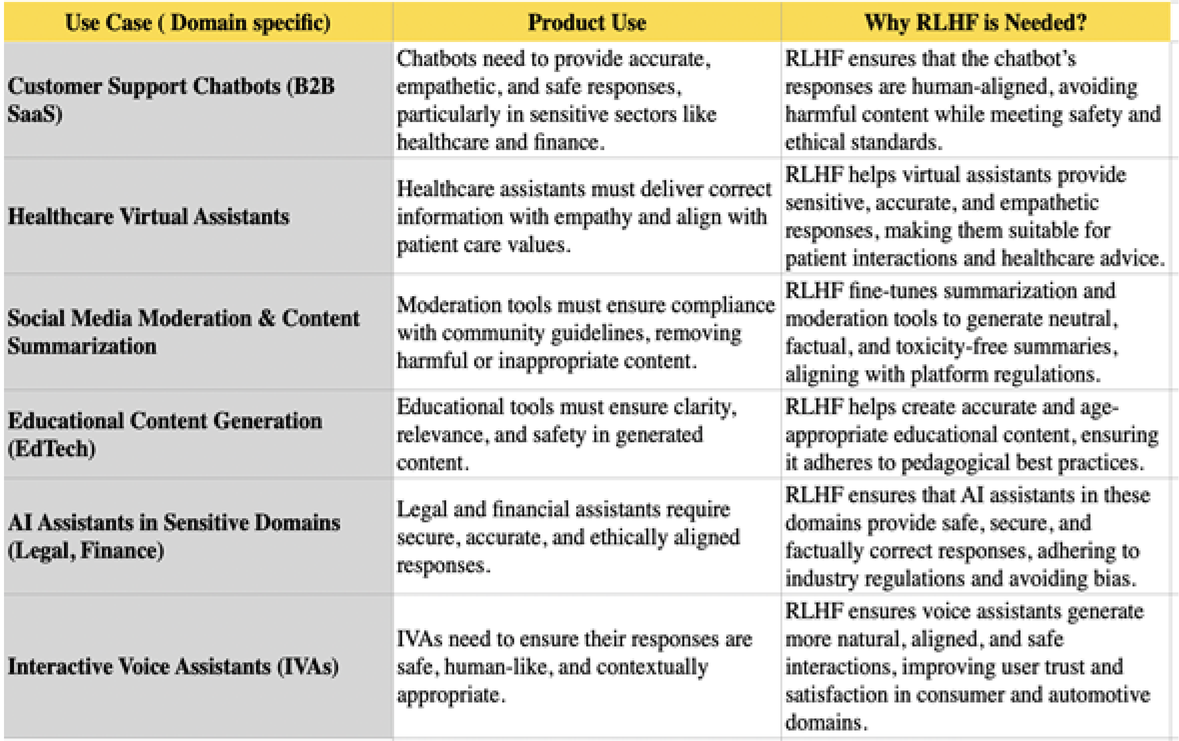
LLM: Just ignore it; it's all in your head.

The LLM responds insensitively, potentially causing harm to the user's well-being.

Enhancing Ethical AI through RLHF in Domain-Specific Use Cases

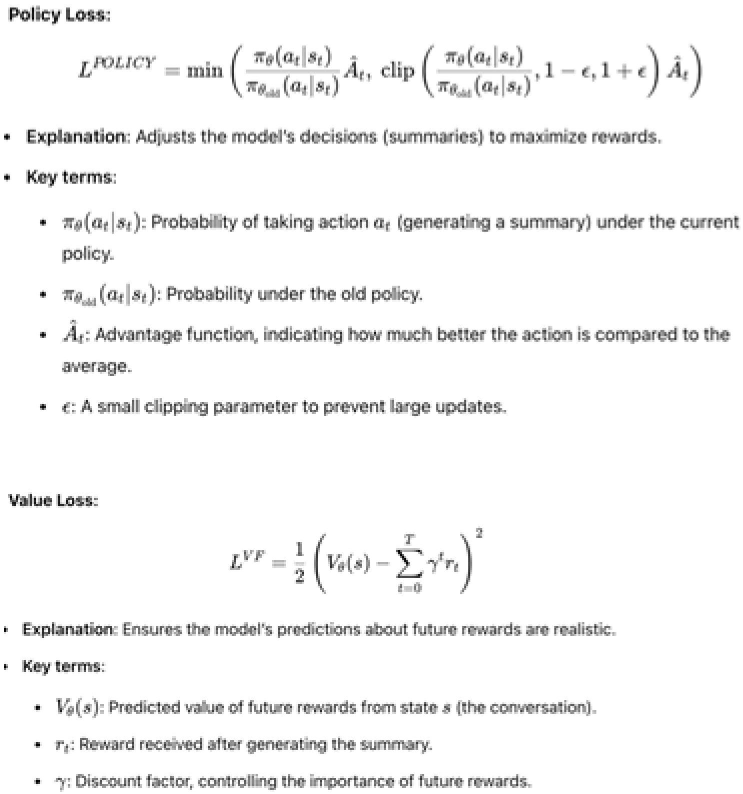
The primary objective of **ChatGPT** (powered by an underlying LLM) was to generate human-like text, which has led to the development of various applications.

Our objective is to refine the model's(LLMs) general capabilities to align with **Ethical AI principles** and make it suitable for more **sensitive use cases**.



Use Case (Domain specific)	Product Use	Why RLHF is Needed?
Customer Support Chatbots (B2B SaaS)	Chatbots need to provide accurate, empathetic, and safe responses, particularly in sensitive sectors like healthcare and finance.	RLHF ensures that the chatbot's responses are human-aligned, avoiding harmful content while meeting safety and ethical standards.
Healthcare Virtual Assistants	Healthcare assistants must deliver correct information with empathy and align with patient care values.	RLHF helps virtual assistants provide sensitive, accurate, and empathetic responses, making them suitable for patient interactions and healthcare advice.
Social Media Moderation & Content Summarization	Moderation tools must ensure compliance with community guidelines, removing harmful or inappropriate content.	RLHF fine-tunes summarization and moderation tools to generate neutral, factual, and toxicity-free summaries, aligning with platform regulations.
Educational Content Generation (EdTech)	Educational tools must ensure clarity, relevance, and safety in generated content.	RLHF helps create accurate and age-appropriate educational content, ensuring it adheres to pedagogical best practices.
AI Assistants in Sensitive Domains (Legal, Finance)	Legal and financial assistants require secure, accurate, and ethically aligned responses.	RLHF ensures that AI assistants in these domains provide safe, secure, and factually correct responses, adhering to industry regulations and avoiding bias.
Interactive Voice Assistants (IVAs)	IVAs need to ensure their responses are safe, human-like, and contextually appropriate.	RLHF ensures voice assistants generate more natural, aligned, and safe interactions, improving user trust and satisfaction in consumer and automotive domains.

Loss Functions Used



Policy Loss:

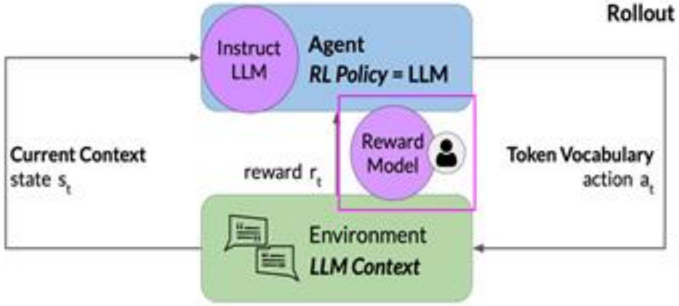
$$L^{POLICY} = \min\left(\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}\hat{A}_t,\ \text{clip}\left(\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}, 1-\epsilon, 1+\epsilon\right)\hat{A}_t\right)$$

- **Explanation**: Adjusts the model's decisions (summaries) to maximize rewards.
- **Key terms**:
  - $\pi_\theta(a_t|s_t)$: Probability of taking action a_t (generating a summary) under the current policy.
  - $\pi_{\theta_{old}}(a_t|s_t)$: Probability under the old policy.
  - $\hat{A}_t$: Advantage function, indicating how much better the action is compared to the average.
  - ϵ: A small clipping parameter to prevent large updates.

Value Loss:

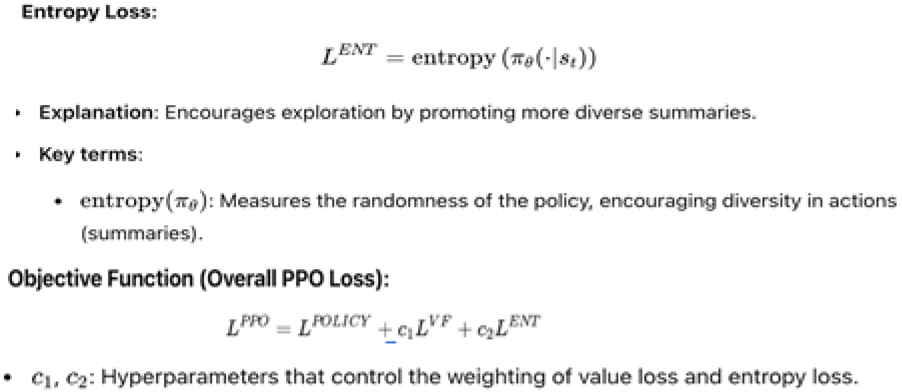
$$L^{VF} = \frac{1}{2}\left(V_\theta(s) - \sum_{t=0}^{T}\gamma^t r_t\right)^2$$

- **Explanation**: Ensures the model's predictions about future rewards are realistic.
- **Key terms**:
  - $V_\theta(s)$: Predicted value of future rewards from state s (the conversation).
  - r_t: Reward received after generating the summary.
  - γ: Discount factor, controlling the importance of future rewards.



conversation text that the LLM summarize

conversation summary



Entropy Loss:

$$L^{ENT} = \text{entropy}\left(\pi_\theta(\cdot|s_t)\right)$$

- **Explanation**: Encourages exploration by promoting more diverse summaries.
- **Key terms**:
  - $\text{entropy}(\pi_\theta)$: Measures the randomness of the policy, encouraging diversity in actions (summaries).

Objective Function (Overall PPO Loss):

$$L^{PPO} = L^{POLICY} + c_1 L^{VF} + c_2 L^{ENT}$$

- c_1, c_2: Hyperparameters that control the weighting of value loss and entropy loss.

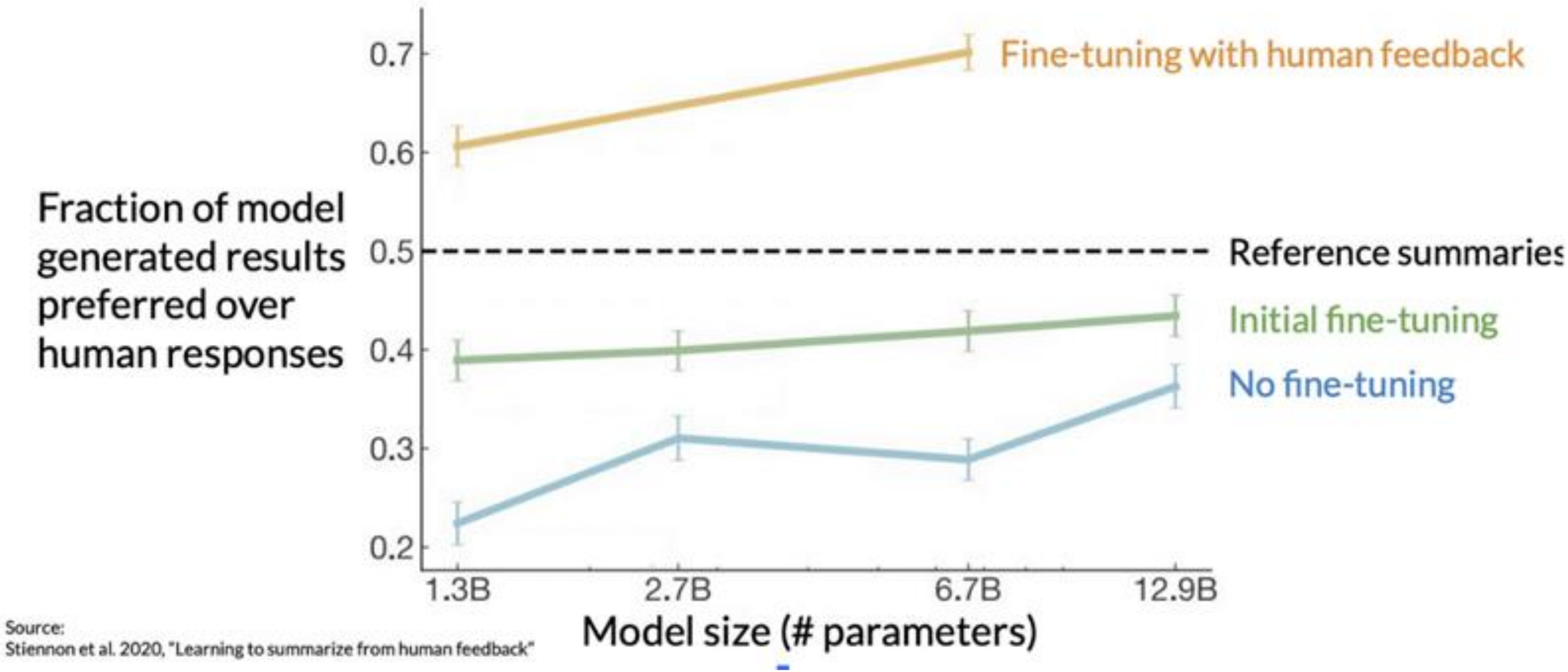
The **objective function** in this context is the **overall PPO loss function** that combines the three loss components: **policy loss**, **value loss**, and **entropy loss**. The objective is to minimize this combined loss, which balances improving the model's actions (policy), making accurate reward predictions (value), and encouraging exploration (entropy).

Image source: DeepLearning.AI

Comparison of Summarization Results: Before and After RLHF Fine-Tuning

Query	Summarize the following conversation. #Person1#: Judy, what is everybody talking about? #Person2#: Haven't you heard? Richard was fired by our manager. #Person1#: You're kidding. It can't be true. #Person2#: Believe it or not. Everybody is talking about it in the company. #Person1#: Really? I'm surprised. #Person2#: Me too. Summary: </s	Summarize the following conversation. #Person1#: Could you help me figure out how to look for a job? #Person2#: We have lots of options, what type of job do you need? #Person1#: I want to work in an office. #Person2#: Do you want to work part-time or full-time? #Person1#: I want to work full-time. #Person2#: We have binders with local job listings or you can make use of the computers. OK? #Person1#: I am confused a bit but I am sure that I can figure it out. #Person2#: If you make an appointment with a job counselor, they can also provide you with a lot of information. Do you want to see a counselor? #Person1#: No, I don't think I need to do that. #Person2#: Well, you know that this job center is here for your use, so help yourself. Good luck! Summary: </s>
Response Before :	Judy learns from a colleague that Richard was fired by their manager, which becomes a hot topic of gossip in the company. They express disbelief and surprise over the news, with everyone talking about it. **Reward : 1.169774**	<pad> #Person2# helps #Person1# find a job, with photos of local job listings and information. #Person1# will see a counselor that can help him easily.</s> **Reward : 2.017895**
Response After	Judy’s colleague informs her that Richard’s departure from the company was decided by their manager. The news comes as a surprise to both of them, and they discuss how it’s a widely discussed topic among their coworkers. **Reward : 1.981302**	<pad> #Person1# wants to work in an office, and #Person2# offers suggestions on the job center. They have several offices where they can find both of their jobs. Then they try to find a job counselor. The counselor can give the information for no time.</s> **Reward :2.370514**

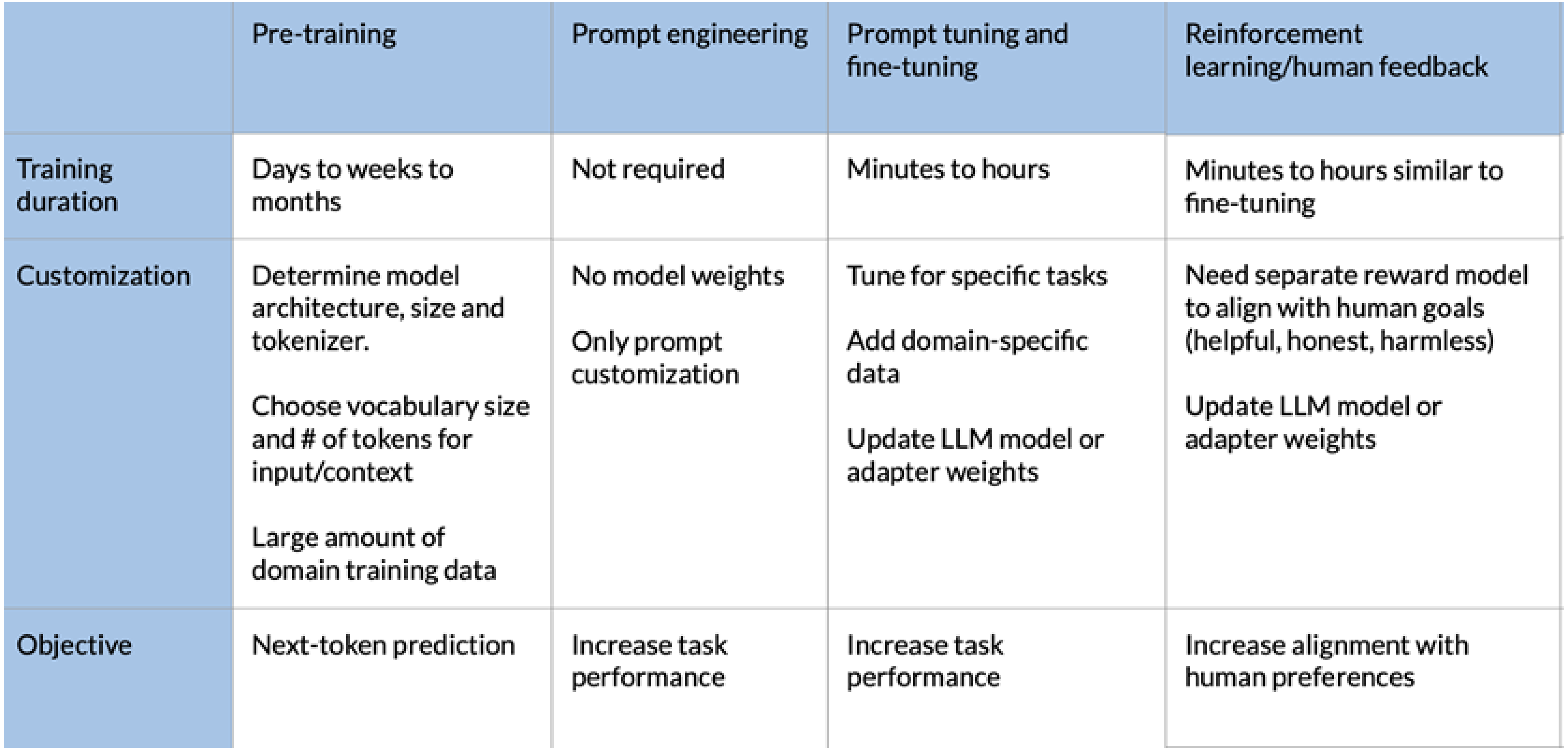
Impact of Fine-Tuning with Human Feedback on Model Performance



Source:
Stiennon et al. 2020, "Learning to summarize from human feedback"

https://openai.com/index/learning-to-summarize-with-human-feedback/

Comparison of Training Methods for Large Language Models (LLMs)

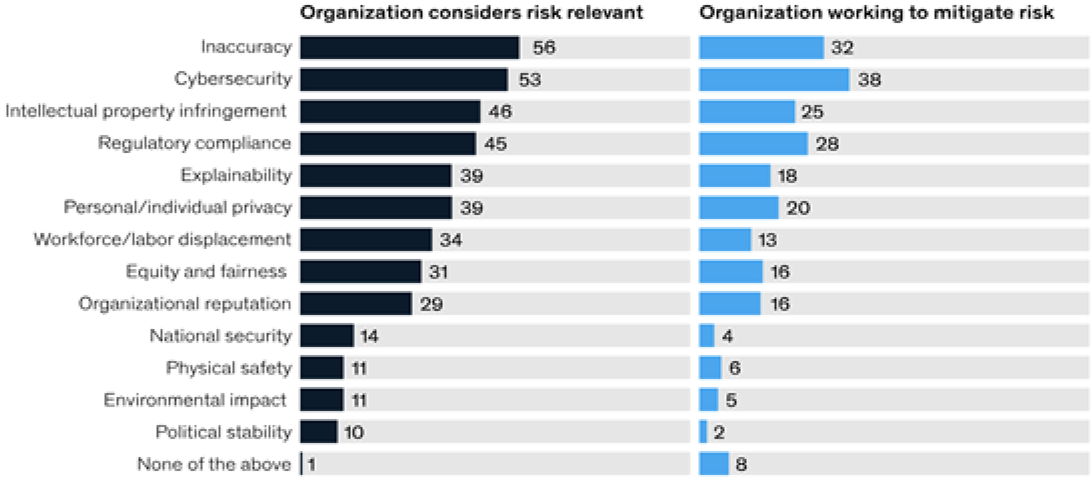


	Pre-training	Prompt engineering	Prompt tuning and fine-tuning	Reinforcement learning/human feedback
Training duration	Days to weeks to months	Not required	Minutes to hours	Minutes to hours similar to fine-tuning
Customization	Determine model architecture, size and tokenizer. Choose vocabulary size and # of tokens for input/context Large amount of domain training data	No model weights Only prompt customization	Tune for specific tasks Add domain-specific data Update LLM model or adapter weights	Need separate reward model to align with human goals (helpful, honest, harmless) Update LLM model or adapter weights
Objective	Next-token prediction	Increase task performance	Increase task performance	Increase alignment with human preferences

Generative AI Risks and RLHF Impact on Industry Applications

Inaccuracy, cybersecurity, and intellectual property infringement are the most-cited risks of generative AI adoption.

Generative AI–related risks that organizations consider relevant and are working to mitigate, % of respondents[1]



[1] Asked only of respondents whose organizations have adopted AI in at least 1 function. For both risks considered relevant and risks mitigated, n = 913.
Source: McKinsey Global Survey on AI, 1,684 participants at all levels of the organization, April 11–21, 2023

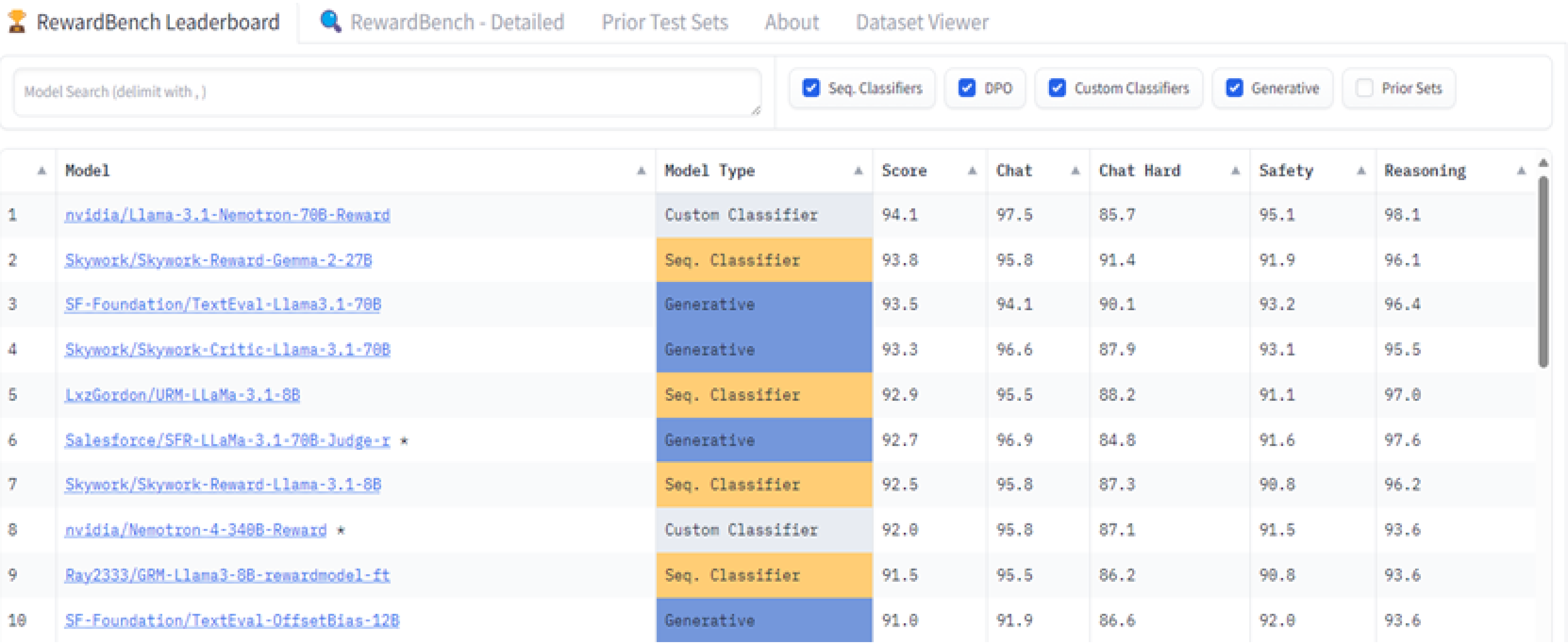
McKinsey & Company

"Keep a human in the loop"
Make sure a human checks any generative AI output before it's published or used.
– mckinsey :

Source: *The state of AI in 2023: Generative AI's breakout year*

Generative AI Risks and RLHF Impact on Industry Applications (cont’d)

BCG highlights that generative AI, driven by RLHF, is transforming industries like **e-commerce** through conversational commerce. This allows AI systems to better guide users through complex purchase processes while ensuring safe, personalized, and human-like interaction, reducing customer service costs by around 30%- bcg

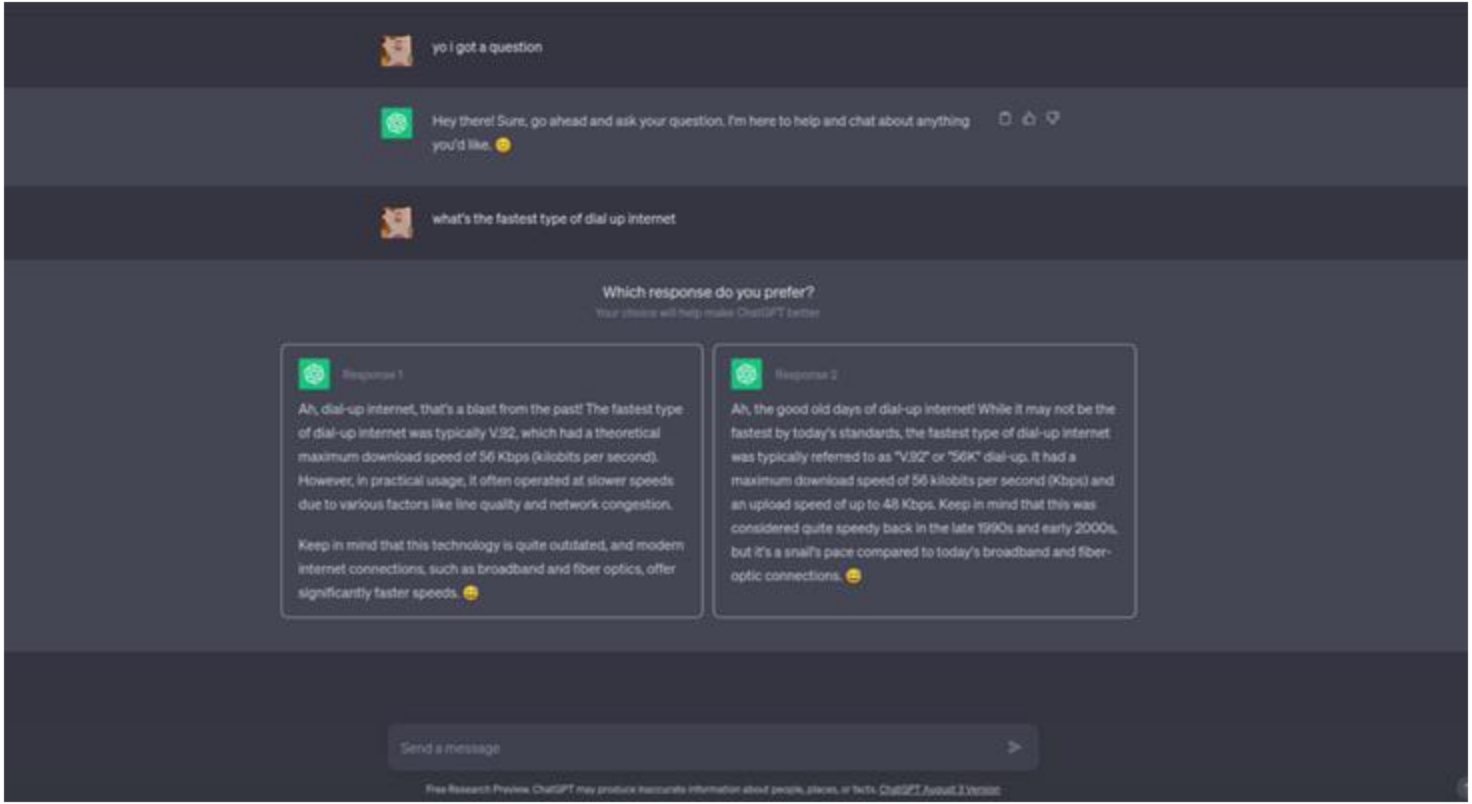


☑ Seq. Classifiers ☑ DPO ☑ Custom Classifiers ☑ Generative □ Prior Sets

	Model	Model Type	Score	Chat	Chat Hard	Safety	Reasoning
1	nvidia/Llama-3.1-Nemotron-70B-Reward	Custom Classifier	94.1	97.5	85.7	95.1	98.1
2	Skywork/Skywork-Reward-Gemma-2-27B	Seq. Classifier	93.8	95.8	91.4	91.9	96.1
3	SF-Foundation/TextEval-Llama3.1-70B	Generative	93.5	94.1	90.1	93.2	96.4
4	Skywork/Skywork-Critic-Llama-3.1-70B	Generative	93.3	96.6	87.9	93.1	95.5
5	LxzGordon/URM-LLaMa-3.1-8B	Seq. Classifier	92.9	95.5	88.2	91.1	97.0
6	Salesforce/SFR-LLaMa-3.1-70B-Judge-r *	Generative	92.7	96.9	84.8	91.6	97.6
7	Skywork/Skywork-Reward-Llama-3.1-8B	Seq. Classifier	92.5	95.8	87.3	90.8	96.2
8	nvidia/Nemotron-4-340B-Reward *	Custom Classifier	92.0	95.8	87.1	91.5	93.6
9	Ray2333/GRM-Llama3-8B-rewardmodel-ft	Seq. Classifier	91.5	95.5	86.2	90.8	93.6
10	SF-Foundation/TextEval-OffsetBias-12B	Generative	91.0	91.9	86.6	92.0	93.6

Nvidia - Reward Model

Future Scope : Direct Preference Optimization (DPO)



yo i got a question

Hey there! Sure, go ahead and ask your question. I'm here to help and chat about anything you'd like. 😊

what's the fastest type of dial up internet

Which response do you prefer?

Your choice will help make ChatGPT better

Response 1

Ah, dial-up internet, that's a blast from the past! The fastest type of dial-up internet was typically V.92, which had a theoretical maximum download speed of 56 Kbps (kilobits per second). However, in practical usage, it often operated at slower speeds due to various factors like line quality and network congestion.

Keep in mind that this technology is quite outdated, and modern internet connections, such as broadband and fiber optics, offer significantly faster speeds. 😄

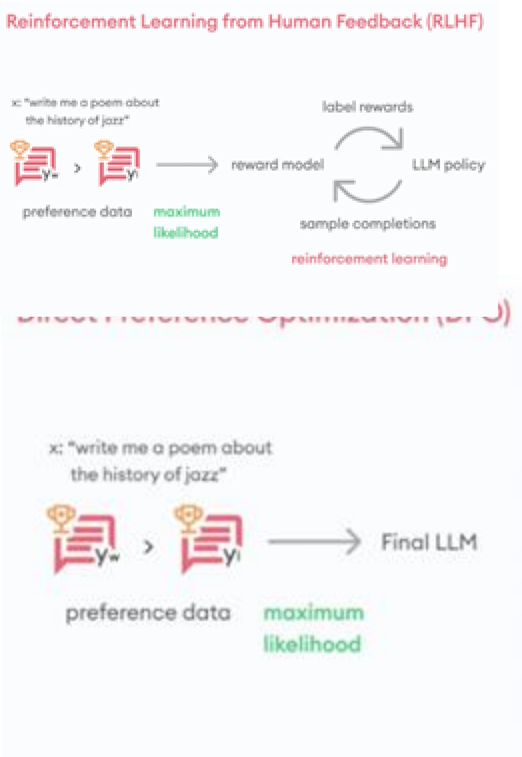
Response 2

Ah, the good old days of dial-up internet! While it may not be the fastest by today's standards, the fastest type of dial-up internet was typically referred to as "V.92" or "56K" dial-up. It had a maximum download speed of 56 kilobits per second (Kbps) and an upload speed of up to 48 Kbps. Keep in mind that this was considered quite speedy back in the late 1990s and early 2000s, but it's a snail's pace compared to today's broadband and fiber-optic connections. 😄

Free Research Preview. ChatGPT may produce inaccurate information about people, places, or facts. ChatGPT August 3 Version

Future Scope : Direct Preference Optimization (DPO)

- Here users provide **direct feedback** by selecting preferred responses from pairs generated by the model. These preferences serve as the optimization signal for model training.
- DPO works by **comparing two model outputs** for the same prompt and using the human preference between them to guide training.
- The core idea is that if one response is preferred over another, the model's parameters should be adjusted so that future outputs are more similar to the preferred response.
- This is done by optimizing the probability distribution of the model, encouraging it to assign higher likelihood to responses that are more aligned with human preferences.
- The optimization process involves using a specially designed loss function (**Binary Cross-Entropy Loss**) that adjusts the model based on these preferences. The model updates are made to increase the probability of the preferred response compared to the non-preferred one.
- One of the key benefits of DPO is its **simplicity** and **efficiency**. Since it avoids the need to train and maintain a separate reward model, DPO reduces the complexity of the fine-tuning pipeline.
- This direct optimization is particularly efficient in scenarios where obtaining human feedback is easier but defining a complex reward model might be challenging or infeasible.



References

Amazon Web Services. (n.d.). *What is reinforcement learning from human feedback?*. *AWS*. https://aws.amazon.com/what-is/reinforcement-learning-from-human-feedback/

DeepLearning.AI. (n.d.). The future of AI and deep learning. *DeepLearning.AI*. https://www.deeplearning.ai/

Dubois, Y., Fawzi, A., & Fawzi, O. (2023). Direct preference optimization: Your language model is secretly a reward model. *arXiv*. https://arxiv.org/abs/2305.18290

IBM. (n.d.). Reinforcement learning from human feedback (RLHF). *IBM*. https://www.ibm.com/topics/rlhf

McKinsey & Company. (2023). What's the future of generative AI? An early view in 15 charts. *McKinsey*. https://www.mckinsey.com/featured-insights/mckinsey-explainers/whats-the-future-of-generative-ai-an-early-view-in-15-charts

NVIDIA Developer. (2023, July 6). *New reward model helps improve LLM alignment with human preferences*. *NVIDIA*. https://developer.nvidia.com/blog/new-reward-model-helps-improve-llm-alignment-with-human-preferences/

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). *Proximal policy optimization algorithms*. *arXiv*. https://arxiv.org/abs/1707.06347

SuperAnnotate. (n.d.). Direct preference optimization (DPO). *SuperAnnotate*. https://www.superannotate.com/blog/direct-preference-optimization-dpo#:~:text=DPO%20changes%20the%20reward%20model,LLMs%20align%20with%20human%20preferences

Thank You

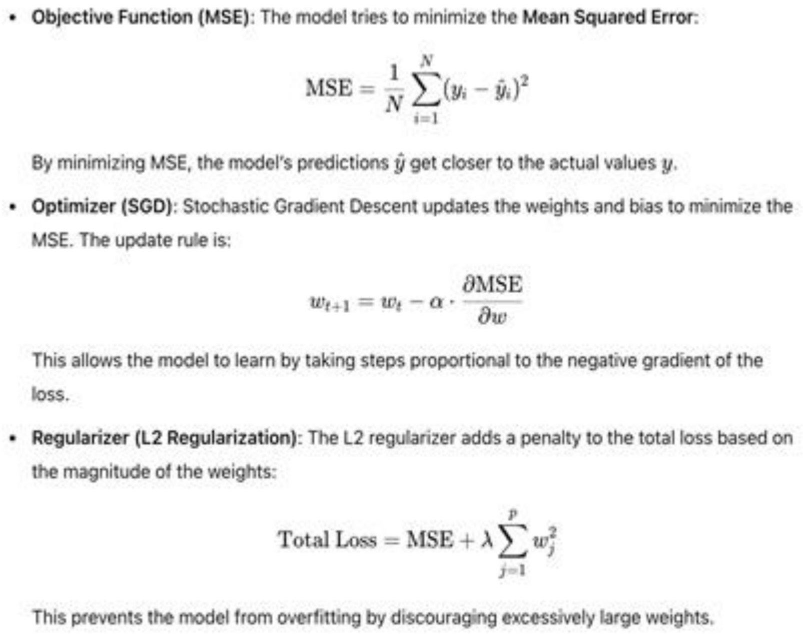
For Your Attention!

Any Questions

APPENDIX

Understanding of Basic DL Training Objectives with Example

- **Objective Function (Loss Function - MSE)**:
  The objective function, typically the loss function, measures how far off the model's predictions are from the actual values. The model's goal is to **minimize this loss** during training to improve its predictions.
- **Optimizer (Adam)**:
  The optimizer is an algorithm that **adjusts the model's parameters** (weights and biases) to minimize the objective function (loss). It updates the parameters iteratively based on the gradients of the loss function.
- **Regularizer (L2 Regularization)**:
  The regularizer adds a **penalty term** to the objective function to discourage large weights, helping prevent overfitting. It encourages simpler models by limiting the size of the parameters, improving generalization to new data.



- **Objective Function (MSE)**: The model tries to minimize the **Mean Squared Error**:

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

By minimizing MSE, the model's predictions $\hat{y}$ get closer to the actual values y.

- **Optimizer (SGD)**: Stochastic Gradient Descent updates the weights and bias to minimize the MSE. The update rule is:

$$w_{t+1} = w_t - \alpha \cdot \frac{\partial \text{MSE}}{\partial w}$$

This allows the model to learn by taking steps proportional to the negative gradient of the loss.

- **Regularizer (L2 Regularization)**: The L2 regularizer adds a penalty to the total loss based on the magnitude of the weights:

$$\text{Total Loss} = \text{MSE} + \lambda \sum_{j=1}^{p} w_j^2$$

This prevents the model from overfitting by discouraging excessively large weights.

With this combination of **MSE (loss function)**, **SGD (optimizer)**, and **L2 regularization (regularizer)**, the model is trained to minimize error, update weights efficiently, and avoid overfitting, leading to better generalization on new data.

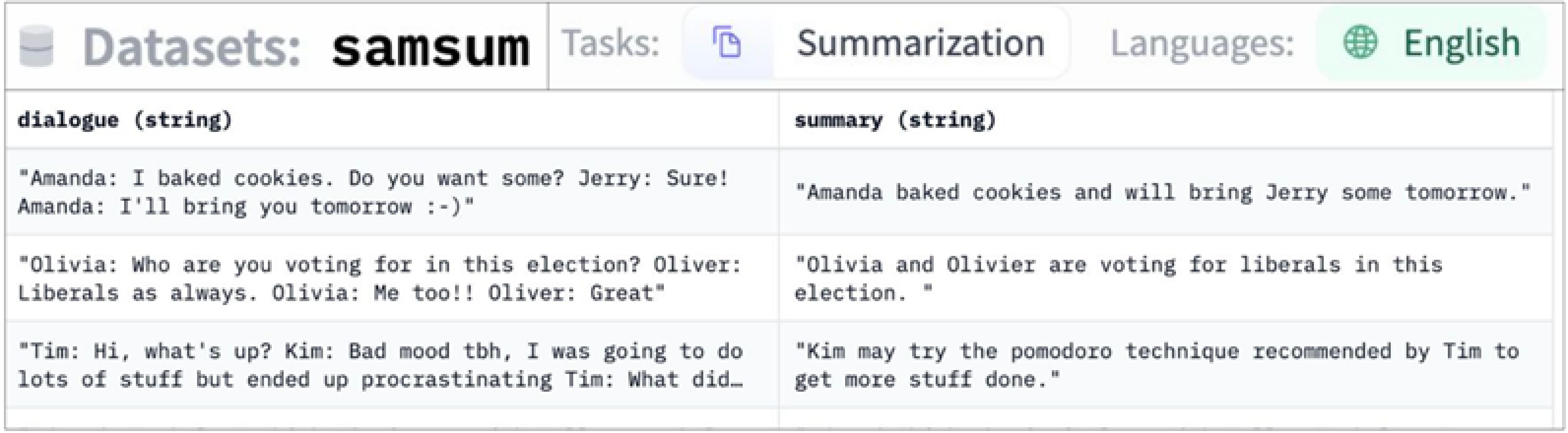
Integrating RLHF with Other Methods

- **RAG vs. RLHF**:
  - RAG improves **factual accuracy** but doesn’t align with **human preferences** or **ethics**.
  - RLHF ensures responses are **empathetic**, **safe**, and **user-friendly**, adding the human preference layer missing in RAG.
- **Instruction Tuning and RLHF**:
  - **Instruction tuning** helps models follow tasks but lacks feedback on **what humans prefer** or find **safer**.
  - RLHF refines the model to improve **user satisfaction** and **response safety**.
- **Agentic AI with RLHF**:
  - Agentic AI enables autonomy but may lack **value alignment** without RLHF.
  - RLHF teaches agentic AI to prioritize **human-centric outcomes**.
- **Integrated Approach**:
  - **Combining RAG and RLHF** ensures both factual accuracy and alignment with **human preferences**.
  - **Instruction tuning** followed by RLHF makes models **task-efficient** and **ethically aligned**.

FLAN-T5 Training Data (an Extension of T5)

SAMSum: A dialogue dataset

Sample prompt training dataset (**samsum**) to fine-tune FLAN-T5 from pretrained T5



Datasets: **samsum** Tasks: Summarization Languages: English

dialogue (string)	summary (string)
"Amanda: I baked cookies. Do you want some? Jerry: Sure! Amanda: I'll bring you tomorrow :-)"	"Amanda baked cookies and will bring Jerry some tomorrow."
"Olivia: Who are you voting for in this election? Oliver: Liberals as always. Olivia: Me too!! Oliver: Great"	"Olivia and Olivier are voting for liberals in this election. "
"Tim: Hi, what's up? Kim: Bad mood tbh, I was going to do lots of stuff but ended up procrastinating Tim: What did…	"Kim may try the pomodoro technique recommended by Tim to get more stuff done."