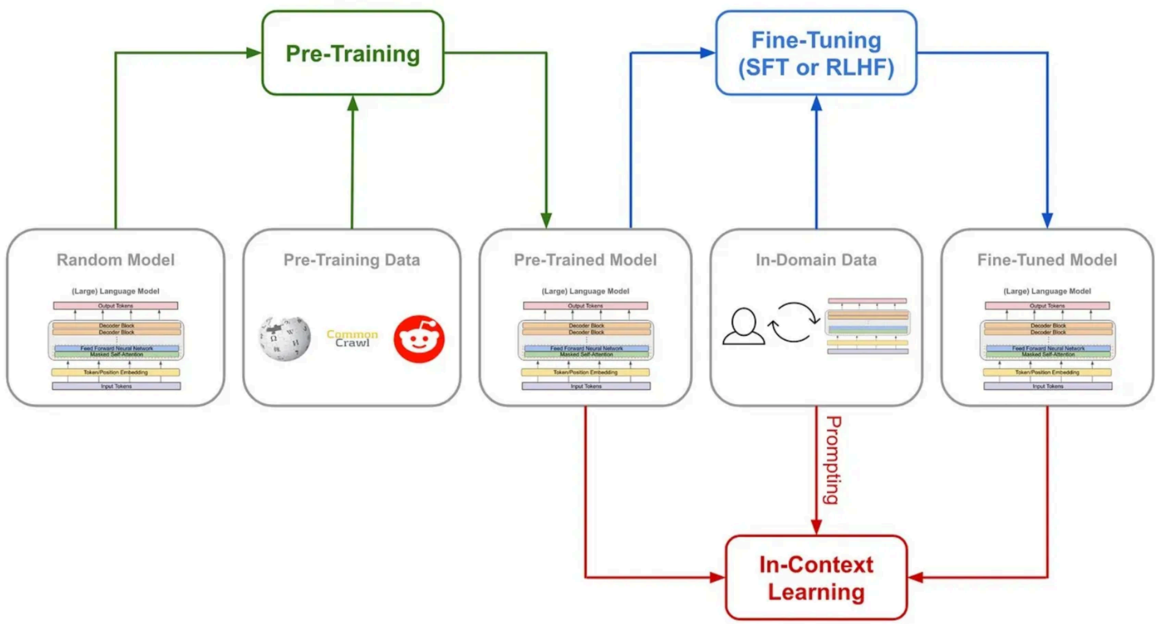
LLM Enterprise Patterns

Iván Moreno (ivan@nieveconsulting.com)

Table of Contents

Fine-tuning Techniques

Supervised Fine-Tuning (SFT)



- Involves training a model on a **labeled dataset** to improve **task-specific** performance.
- Updates all model parameters, which can be in the *billions*. Commonly used for adapting models to significantly different tasks.
- Can potentially lead to **catastrophic forgetting**: forgetting general knowledge learned during pre-training.
- **Examples**: Instruction following, sentiment analysis, named entity recognition.

Parameter-Efficient Fine-Tuning (PEFT)

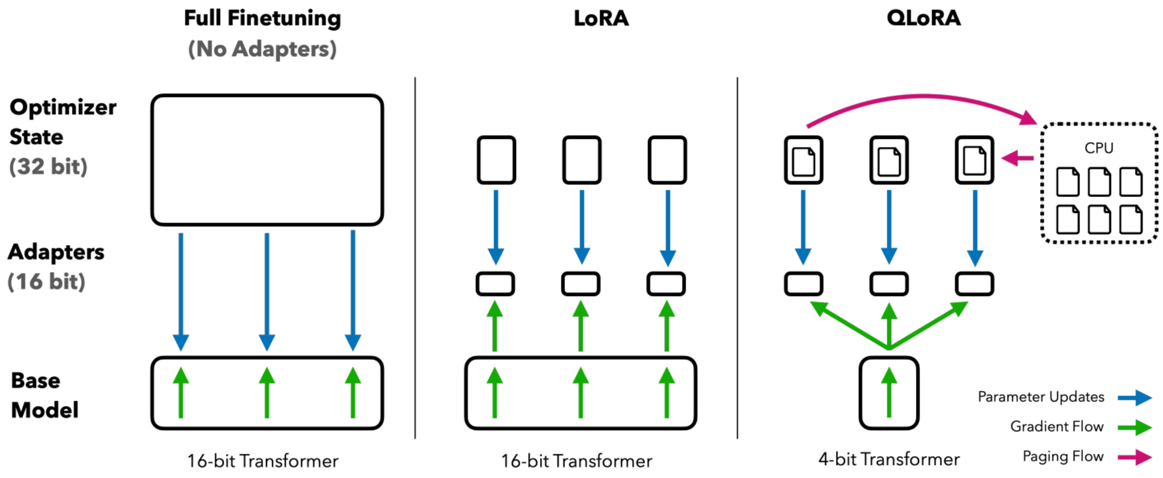


Figure 1: Different finetuning methods and their memory requirements. QLoRA improves over LoRA by quantizing the transformer model to 4-bit precision and using paged optimizers to handle memory spikes.

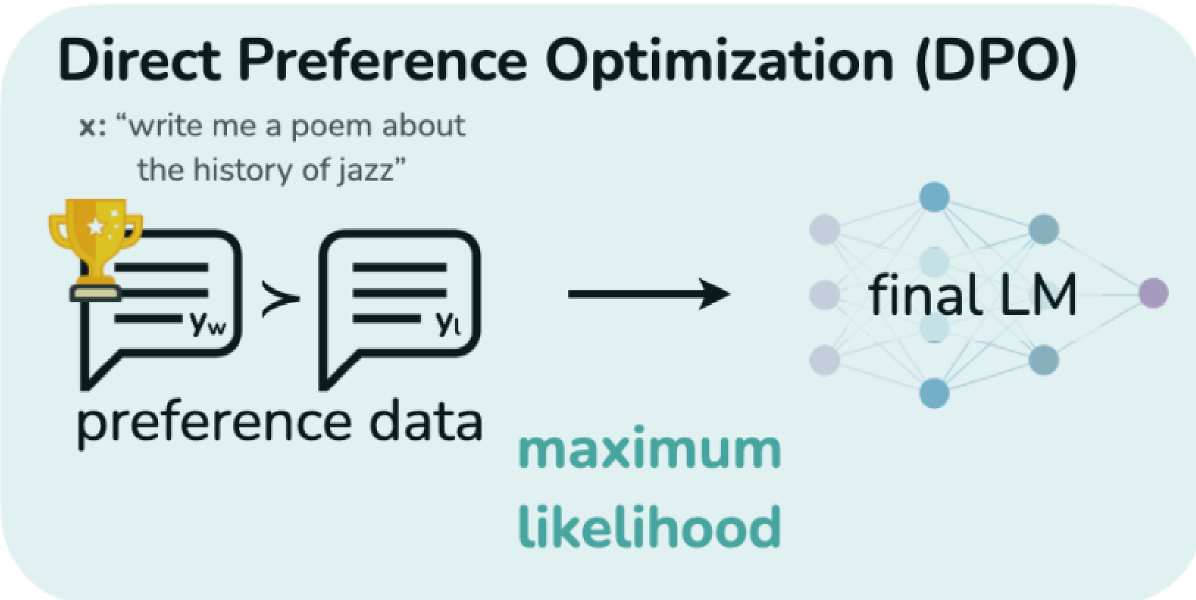
- Focuses on updating a **small subset** of model parameters.
- Often involves *less than 1%* of the total parameters, or adding a small number of new parameters (adapter layers).
- It requires significantly less data and computational resources compared to SFT.
- *Retains* the general-purpose knowledge learned during pre-training.
- **LoRa (Low-Rank Adaptation)**: A method for parameter-efficient fine-tuning that uses low-rank matrices to adapt models to new tasks.

Alignment Techniques

- **Definition**: Ensuring that LLMs generate outputs **aligned** with human preferences and expectations.
- **Common Methods**: Reinforcement Learning with Human AI Feedback (RLHF), Direct Preference Optimization (DPO).

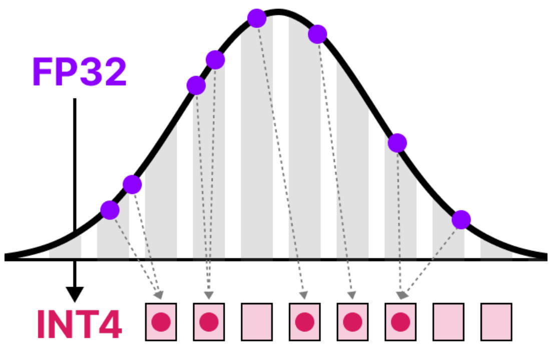
Reinforcement Learning with Human AI Feedback (RLHF)

Direct Preference Optimization (DPO)



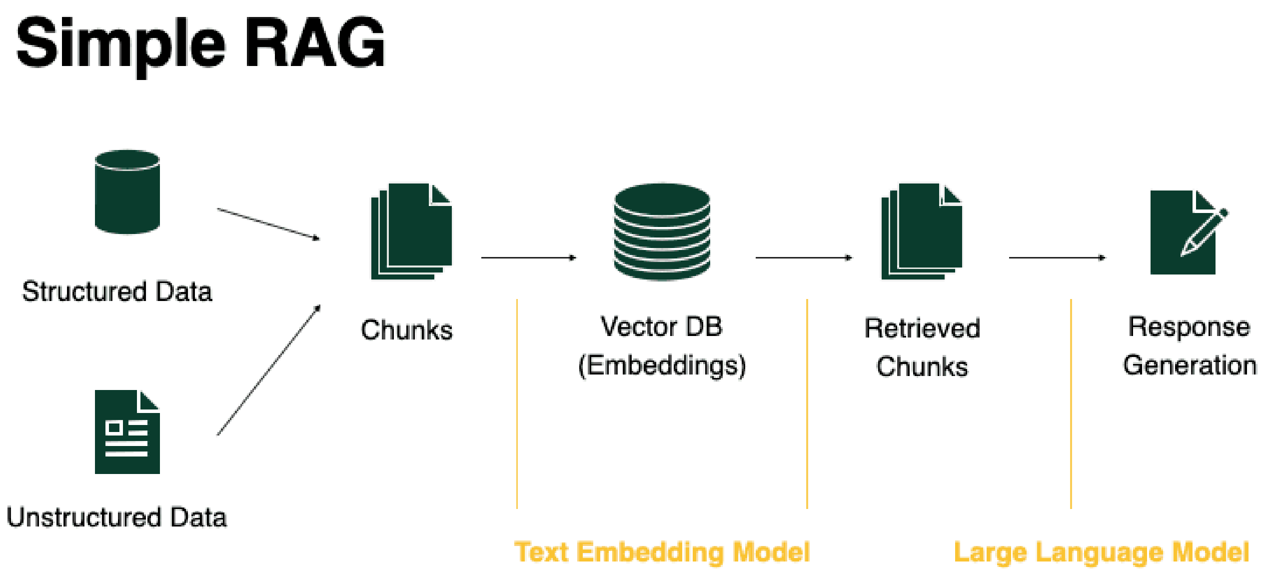
- **Definition:** Optimizing LLMs directly using **pairwise preference data** instead of RL.
- **Process:** Training models to predict preferences between pairs of outputs.
- **Example input:**
  - **Q:** How's going to be the next president of the USA?
  - **Chosen:** As a language model, I can't predict the future.
  - **Rejected:** <CANDIDATE> is going to win the election.

Quantization



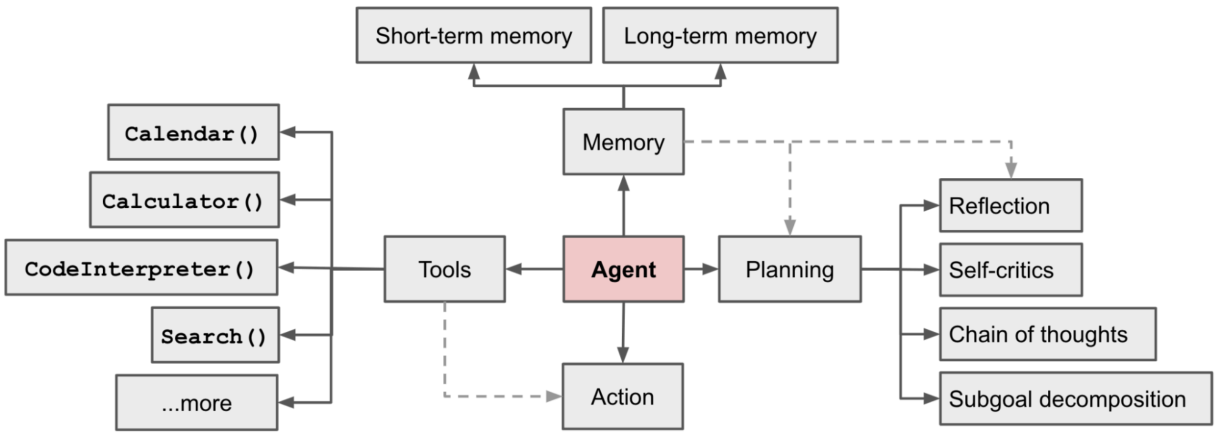
- **Reducing the precision** of model weights to lower inference costs without significant accuracy loss.
- Comes at the cost of **reduced model performance**. The degree of performance loss depends on the quantization level.
- **Types**:
  - **8-bit Quantization:** Most commonly used. Reduces the precision of weights to 8 bits.
  - **4-bit Quantization:** More aggressive. Reduces the precision of weights to 4 bits.
  - **Mixed Precision Quantization:** Combines 8-bit and 4-bit quantization to optimize performance, depending on each layer's sensitivity to precision loss.

Retrieval-Augmented Generation



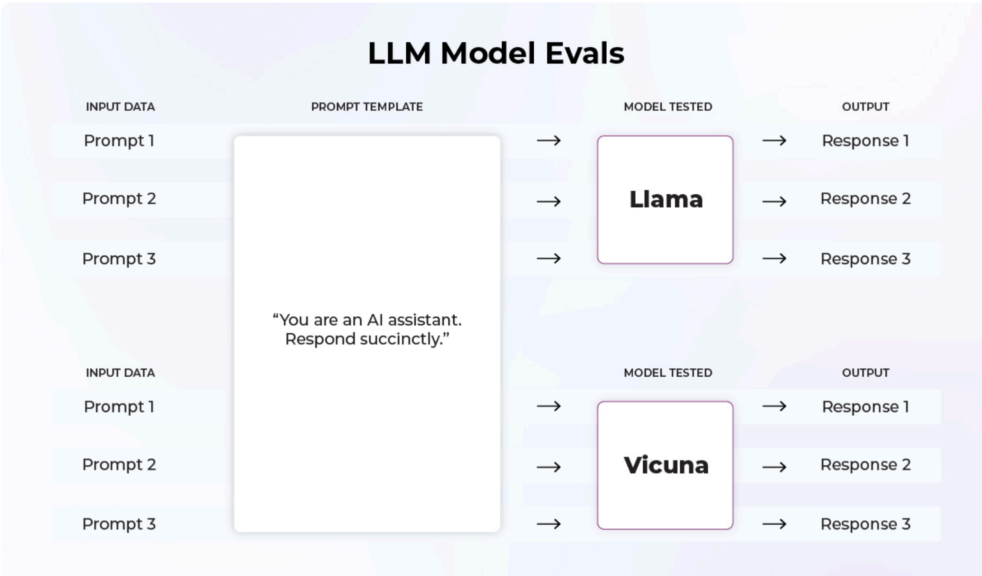
- **Definition**: Combining retrieval systems with LLMs for better response generation.
- LLMs suffer from **fixed knowledge** at training time, while retrieval systems can provide up-to-date information.
- By leveraging additional context, we reduce the potential for **hallucinations**.
  - **Retriever**: Retrieves relevant information from a knowledge base.
  - **Generator**: Generates responses based on the retrieved information.

LLM Agents & Agentic Systems



- **Definition**: Systems that allow LLMs to **autonomously perform tasks** by interacting with APIs, databases, or other tools.
- **Example**: A support ticket system that uses an LLM to generate responses and interact with a CRM.
- **Components**:
  - **Agent**: The LLM breaks down tasks into subtasks and interacts with external tools to complete them.
  - **Tool**: External APIs, databases, or other tools that the agent interacts with.

LLM Evaluation

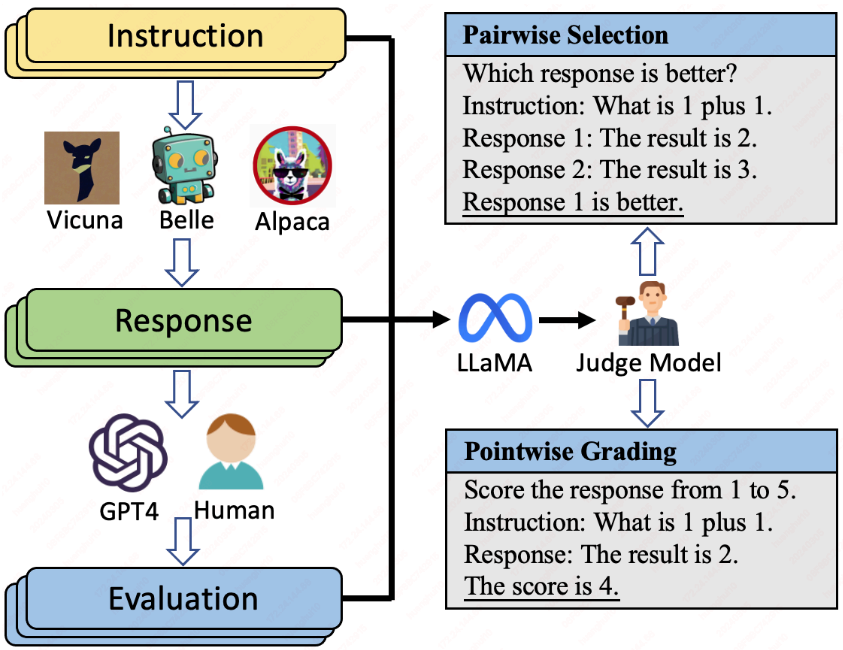


- The challenge of evaluating LLM outputs is that is **subjective** and depends on the task.
- We can **quantitatively** evaluate LLM outputs if we have a **labeled dataset**, but this is *not always possible*.
- Additionally, the challenge is how to evaluate the **qualitative aspects** of the output, such as coherence, tone, and relevance.
- Human evaluation is still the most reliable method for evaluation when it comes to these qualitative aspects.

Common Metrics

- **Perplexity**: inverse of the probability that the model assigns to a sequence of words, normalized by the number of words.
  - A higher perplexity indicates lower model performance, as the model is *“less certain”* about the sequence.
- **BLEU**: Evaluates the overlap between predicted and reference text.
  - Uses *n-grams* to measure the *similarity* between the predicted and reference text.
  - Focuses on the number of overlapping n-grams, which can be *problematic* for long sequences.
- **ROUGE**: Measures **recall** for sequence generation tasks.
  - There are several *variants* of ROUGE:
    - *ROUGE-N*: Measures n-gram overlap (e.g., ROUGE-1 measures unigram overlap).
    - ROUGE-L: Measures the longest common subsequence between the predicted and reference text.
    - *ROUGE-W*: Measures the weighted longest common subsequence, giving more weight to longer subsequences.
    - *ROUGE-S*: Measures skip-bigram overlap, which allows for gaps between words.
  - Higher ROUGE scores indicate better performance (e.g., ROUGE-2 measures bigra

LLM-as-Judge


Instruction
Vicuna
Belle
Alpaca
Response
GPT4
Human
Evaluation
LLaMA
Judge Model
Pairwise Selection
Which response is better?
Instruction: What is 1 plus 1.
Response 1: The result is 2.
Response 2: The result is 3.
Response 1 is better.
Pointwise Grading
Score the response from 1 to 5.
Instruction: What is 1 plus 1.
Response: The result is 2.
The score is 4.

- **Definition**: Using LLMs to **judge** their own or other model outputs based on predefined criteria.
- **Process**: The LLM generates a score based on the predefined **criteria** and provides feedback to the user.
- Papers have *quantified the alignment* between human and LLM judgments, showing that LLMs can be effective judges.
- Some LLMs can be fine-tuned to perform this task more effectively.