Structured summary of the provided section (Post-Training: SFT and Preference Finetuning)

---

1. Big Picture: Post-Training Pipeline

Modern foundation models are typically trained in three stages:

1. Self-supervised pre-training
   - Trained on massive internet-scale data.
   - Learns language structure and broad knowledge.
   - Behavior is general and not aligned for safe or helpful conversations.
2. Supervised Fine-Tuning (SFT)
   - Trained on high-quality (prompt, response) demonstration data.
   - Teaches the model how to behave in conversations.
   - Also called behavior cloning.
3. Preference Finetuning (e.g., RLHF, DPO)
   - Aligns the model with human preferences.
   - Makes responses safer, more helpful, and customer-appropriate.

This combination is common but not mandatory; some systems skip certain steps.

---

2. Supervised Fine-Tuning (SFT)

Why SFT is needed

A pre-trained model is optimized for text completion, not dialogue.
Example:
Input: "How to make pizza"
The model may:

- Add context
- Ask follow-up questions
- Or provide instructions

If the goal is helpful assistance, the third is correct. SFT teaches this behavior explicitly.

How SFT works

- Use demonstration data: (prompt, response) pairs.
- Labelers create high-quality responses.

- The model learns to imitate these examples.

Characteristics of SFT Data

- Covers diverse tasks: Q&A, summarization, translation, explanation, etc.
- Requires skilled annotators.
- Expensive and time-consuming.
- Quality and demographic bias of labelers affect model behavior.

Alternatives

- Volunteer-generated data (e.g., community datasets)
- Heuristic filtering of internet dialogues
- Synthetic (AI-generated) data
- Training from scratch on instruction data (less effective than pre-training + finetuning)

---

3. Preference Finetuning

Why SFT is not enough

SFT teaches *how* to respond, not *what values to follow*.
It does not solve:

- Harmful instructions (e.g., illegal activities)
- Controversial political or social questions
- Value conflicts across cultures

Preference finetuning attempts to align the model with human preferences.

---

4. RLHF (Reinforcement Learning from Human Feedback)

RLHF consists of two stages:

Step 1: Train a Reward Model (RM)

The reward model:

- Takes (prompt, response)
- Outputs a scalar score indicating quality

Instead of giving absolute scores (pointwise evaluation), labelers compare responses:

Format:
(prompt, winning_response, losing_response)

This is easier and more consistent than direct scoring.

Reward Model Objective

For:

- x = prompt
- yw = winning response
- yl = losing response

The model is trained to maximize:

$r(x, yw) - r(x, yl)$

Using a sigmoid-based loss:

$-\log(\sigma(r(x, yw) - r(x, yl)))$

Goal: assign higher scores to preferred responses.

Step 2: Optimize the Base Model

Once the reward model is trained:

1. Sample prompts.
2. Generate responses from the model.
3. Score responses with the reward model.
4. Update the model to maximize reward scores.

This is typically done using PPO (Proximal Policy Optimization), a reinforcement learning algorithm.

---

5. DPO vs RLHF

- RLHF: More complex, flexible, involves reinforcement learning.
- DPO (Direct Preference Optimization): Simpler alternative.
- Some organizations (e.g., Meta for newer models) have moved from RLHF to DPO to reduce complexity.
- Debate continues about why these methods work so well.

---

6. Best-of-N Strategy (Without RL)

Some companies skip reinforcement learning entirely.

Instead:

1. Generate N outputs.
2. Score them with the reward model.
3. Select the highest-scoring output.

This “best-of-N” approach leverages sampling diversity without full RL optimization.

---

7. Key Conceptual Insight

- Pre-training creates a powerful but unaligned model.
- SFT makes it conversational.
- Preference finetuning aligns it with human expectations.

Both SFT and preference finetuning compensate for the noisy and uncontrolled nature of internet-scale pre-training data.

If pre-training data or methods improve dramatically in the future, these post-training steps might become less necessary.
