9 LLM Odds-and-Ends; Chain-of-Thought & Reasoning

(need to add `https://www.datocms-assets.com/64837/1763496622-dr_tulu_draft.pdf` *for more tool use.)* ☹ 10/25

Announcements.
1. Final projects.
2. hw grades.

Plan for today.
1. Two training phases.
2. Alignment.
3. Tool use.
4. Reasoning.

9.1 Two LLM training phases.

1. **Two phases.** Training is typically divided into *pre-training* and *post-training*.
2. **Pre-training.** next-token prediction on all prefixes of documents from a large, curated, *reweighted* corpus. The reweighting, meaning that data from different sources have their sampling frequencies reweighted, are interesting and constantly changing. Also, as of 2025, there is quite a bit of purely machine-labeled data (e.g., for tool use).
3. **Post-training.** SFT or RL on extremely targeted data, with (at least?!) four goals:
   - Instruction following: without this, the model just continues a stream of thought, and doesn't know to explicitly answer questions and follow instructions (Ouyang et al., 2022; Taori et al., 2023). Getting this right is a delicate engineering challenge and it's remarkable how well modern models can follow and remember instructions as we flood their context.
   - Alignment: fit more specifically with human tastes, preferences, ethics, safety, etc (Christiano et al., 2017; Rafailov et al., 2023).
   - Reasoning, specifically with chain-of-thought: this is the ability of the models to essentially pause and think silently, we'll discuss it in detail below.
   - Tool use (the technology behind "agentic" models): the ability of LLMs to invoke external processes and receive tokens from them; also discussed in detail below.

   Remark: All of this is a moving target. For instance, the mid-2025 deepseek 3.2exp paper (DeepSeek-AI, 2025b), states it only uses one phase which simultaneously balances reasoning, tool use, and alignment:

   > *Unlike in previous DeepSeek models, which are trained with multi-stage reinforcement learning, we merge reasoning, agent, and human alignment training into one RL stage. This approach effectively balances performance across diverse domains while circumventing the catastrophic forgetting issues commonly associated with multi-stage training paradigms.*

4. **Methods.** These three post-training goals are, methodologically, approached as follows. Throughout, we will call the pre-trained model the *base model*.
   - SFT: this is simply cross-entropy training on very targeted token sequences. Procedurally it looks like pre-training, but it is given a strong base model and has a relatively tiny dataset.
   - RL: this seems to come in two flavors:

  - Outcome-based rewards, aka the deepseek-RO recipe (DeepSeek-AI, 2025a). This first strategy uses RL with only an extremely clear reward model: marking correct, correctly-formatted answers as positive reward, and all else as 0; there is no effort to evaluate ambiguous answers and award partial credit.
    This is a moving target but for now, this simple no-partial-credit scheme is used the most extensively.
  - Alignment and safety rewards based on human feedback (Christiano et al., 2017; Rafailov et al., 2023).
- Pre-training, specifically the interaction of pre-training with desired outcomes from post-training.
  There is growing understanding that many claimed successes of post-training are now attributed to pre-training (F. Chen et al., 2025). As a concrete example, the "aha" discovered by RL in the deepseek R1 paper (DeepSeek-AI, 2025a) is now largely held to be a consequence of its appearance in pre-training data.

  As such, modern pre-training contains a large amount of reasoning and tool use data, the tool use data being extensive traces that are generated by models which gained their original tool use capabilities purely in post-training.

5. **Instruction tuning.**
   - Without this phase, LLMs simply continue a stream of consciousness. Essential to modern good performance (Ouyang et al., 2022; Taori et al., 2023).
   - Unclear how much this is now explicitly post-trained for, and how much it has been subsumed into alignment/reasoning/tool training.

9.2 Alignment.

1. **Problem description and goals.** The pre-training corpus contains information following some distribution. It is useful and convenient to maintain that data distribution, but often, people then want to "reshape" this distribution after pre-training. This is the task of *alignment*.

   Alignment basically tries to reshape the LLM's output distribution with some human opinions, often according to ideologies, safety, and so on. For example, the pre-training data contains weapon construction, political manifestos, violent fantasies, and so on; a standard use of alignment is to de-emphasize or reverse the perspectives in works of this type.

   Remark. Normally it is presented as increasing safety and other universally good things, but it's important to remember that ultimately some ideology is being baked in, and other people would bake in a different ideology, and ultimately the machine has no choice in the matter.

2. **Preference data and the Bradley-Terry model.** Psychometrics studies have found that it is harder for humans to assign numeric quality scores than it is for them to provide pairwise comparisons. As such, the original alignment papers, indeed the original *RLHF* paper (Christiano et al., 2017), but also the later *DPO* paper (Rafailov et al., 2023), worked with human preference data: given an input or prompt x, let y^+ and y^- respectively denote a pair of preferred and disliked outputs. Supposing some numeric quality score or reward $r(y|x)$, these papers also assumed a standard *Bradley-Terry preference model*, which states that the probability of a human preferring y^+ to y^-, denoted $y^+ \succ y^-$, is given by the difference of the underlying unobserved reward model:

$$\Pr[y^+ \succ y^- \mid x] = \sigma\big(r(y^+|x) - r(y^-|x)\big), \qquad \text{where } \sigma(z) = \frac{1}{1+e^{-z}}.$$

   Methods. There are two primary methods: RLHF, and DPO. What is used in practice is unclear, see the comments in the respective sections below.

   Remarks.