15 Synthetic Data & Distillation

Reinforcement learning from *human feedback* is deeply rooted in the idea of keeping a human influence on the models we are building. When the first models were trained successfully with RLHF, human data was *the only* viable way to improve the models in this way.

Humans were the only way to create high enough quality responses to questions to train on them. Humans were the only way to collect reliable and specific feedback data to train reward models.

As AI models got better, this assumption rapidly broke down. The possibility of synthetic data, which is far cheaper and easier to iterate on, enabled the proliferation from RLHF being the center of attention to the idea of a broader "post-training" shaping the models.

Many reports have been made on how synthetic data causes "model collapse" or other issues in models [197], but this has been emphatically rebuked in leading language models [198] [199]. Synthetic data *can* cause models to have performance issues, but this is caused by using repetitive data or solely data outputted by the model being trained (narrowing its potential distribution) rather than well-rounded data sources.

The leading models **need synthetic data** to reach the best performance. Synthetic data in modern post-training encompasses many pieces of training – language models are used to generate new training prompts from seed examples [200], modify existing prompts, generate completions to prompts [201], provide AI feedback to create preference data [22], filter completions [202], and much more. Synthetic data is key to post-training.

The ability for synthetic data to be impactful to this extent emerged with GPT-4 class models. With early language models, such as Llama 2 and GPT-3.5-Turbo, the models were not reliable enough in generating or supervising data pipelines. Within 1-2 years, language models were far superior to humans for generating answers. In the transition from GPT-3.5 to GPT-4 class models, the ability for models to perform LLM-as-a-judge tasks also emerged. GPT-4 or better models are far more robust and consistent in generating feedback or scores with respect to a piece of content.

Since this transition, the role of synthetic data has only grown in language model training. Otherwise, there are two clear areas where human data continues to be important.

1. The role of human data continues to be at the fringe of capabilities in models – humans must generate data where AI's do not yet have any ability. Once the first strong model exists, synthetic data proliferates.
2. Human preference data is still used in the leading models, even though academic work shows synthetic versions to perform just as well. The role of human preferences is still being established in the literature.

The term distillation has been the most powerful form of discussion around the role of synthetic data in language models. Distillation as a term comes from a technical definition of teacher-student knowledge distillation from the deep learning literature [50].

Distillation colloquially refers to using the outputs from a stronger model to train a smaller model. In post-training, this general notion of distillation takes two common forms:

1. As a data engine to use across wide swaths of the post-training process: Completions for instructions, preference data (or Constitutional AI), or verification for RL.

2. To transfer specific skills from a stronger model to a weaker model, which is often done for specific skill such as mathematic reasoning or coding.

The first strategy has grown in popularity as language models evolved to be more reliable than humans at writing answers to a variety of tasks. GPT-4 class models expanded the scope of this to use distillation of stronger models for complex tasks such as math and code (as mentioned above). Here, distillation motivates having a model suite where often a laboratory will train a large internal model, such as Claude Opus or Gemini Ultra, which is not released publicly and just used internally to make stronger models. With open models, common practice is to distill training data from closed API models into smaller, openly available weights [20]. Within this, curating high-quality prompts and filtering responses from the teacher model is crucial to maximize performance.

Transferring specific skills into smaller language models uses the same principles of distillation – get the best data possible for training. Here, many papers have studying using limited datasets from stronger models to improve alignment [12], mathematic reasoning [203] [204], and test-time scaling [195].