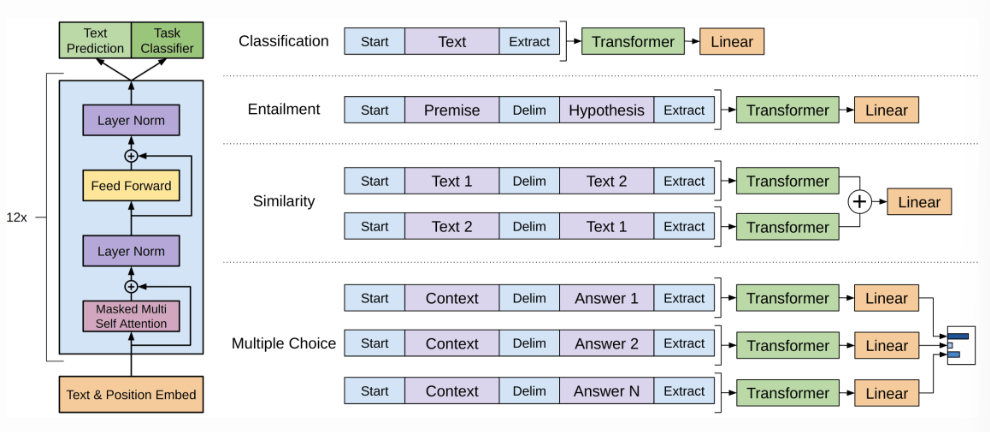
**GPT 3.5 Architecture**

**GPT-n overview**

GPT-n is a language model that operates from left to right, utilizing only the decoder portion of the transformer architecture, without the cross-attention layer. The process begins with word embeddings and positional encodings. Attention mechanisms are then applied to these initial inputs. Following this, several feed-forward layers are used, with regularization steps incorporated at the end. This entire process is encapsulated within a 12-layer transformer decoder. The model is trained using the standard method of cross-entropy loss.

**Fine-tuning**

The fine-tuning loss is composed of two parts: the loss specific to the task at hand, and the loss from language modeling. During the fine-tuning phase, the model’s structure remains unchanged, with the exception of the last linear layer. The input format varies depending on the task. Our focus is on single sentence classification. To classify individual sentences, the data is fed into the model as it was during training, and the label is predicted based on the final representation of the last input token.



**The revolution of GPT-3**

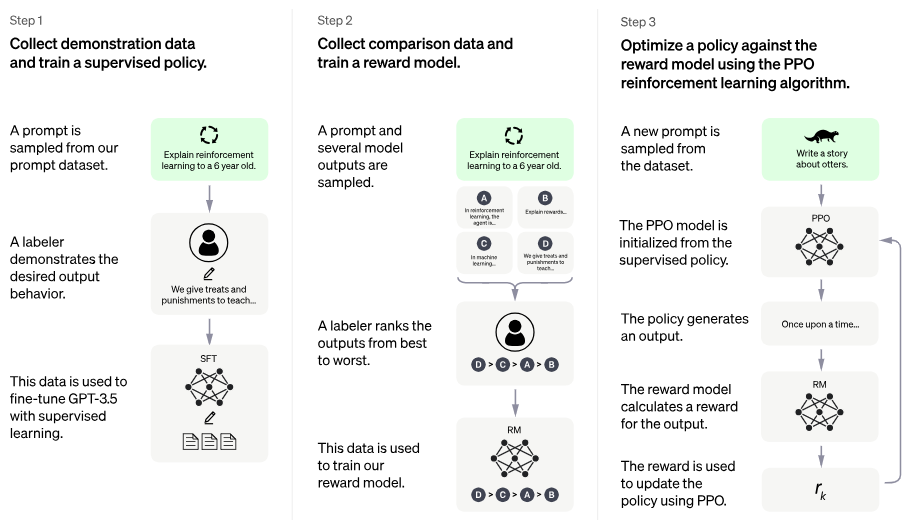
The GPT-3 paper presents a novel concept known as in-context learning, which deviates from traditional artificial intelligence practices. In the conventional supervision paradigm, if we want to classify text as a binary problem, we first create a dataset of positive and negative examples and then train a custom model to make the binary distinction. While this model can be powerful, it may not scale to the complexity of human experience.

In-context learning promises that a single, large, frozen language model can serve all these purposes. We provide the model with examples of positive and negative instances, hoping that it will learn in-context about the distinction we’re trying to make.

The types of n-shot learning that can be introduced to the GPT-3 model include:

* Zero-shot learning: The model can predict the answer given only the task name with no examples.
* One-shot learning: In addition to the task name and description, we provide the model with one example, and the model will be able to predict the answer.
* Few-shot learning: A few examples are introduced to the model along with the task description.

Another significant innovation is the concept of “self-supervision”, a powerful mechanism for acquiring rich representations of form and meaning. In self-supervision, the model’s objective is to learn co-occurrence patterns in the sequences it is trained on. The model is simply learning to assign high probability to attested sequences. These models are thought of as generators, but the generation involves sampling from the model. This is powerful because self-supervision requires minimal human effort and has facilitated the rise of another important mechanism, large-scale pretraining.



The final component is the role of human feedback. The best models, referred to as Instruct models by OpenAI, are trained with more than just self-supervision. Two important aspects from the ChatGPT blog post diagram include:

* The language model is fine-tuned on human-level supervision (Step 2), making binary distinctions about good and bad generations.
* In a second phase, the model generates outputs and humans rank all of the outputs the model has produced (Step 3). This feedback goes into a lightweight reinforcement learning mechanism.

In both of these phases, there are important human contributions that take us beyond the self-supervision step.