Explore the distinctions between prefix language modelling (conditional language modelling) and autoregressive language modelling and describe in short paragraphs.

Prefix language modelling, also known as conditional language modelling, is a technique aimed at fine-tuning large language models for customised natural language generation tasks. It involves keeping the weights of the parameters of a pretrained model fixed while training a small continuous prefix (prompt or input) to the model. Typically, only a small fraction (around 0.1%) of the parameters need to be adjusted. This method enables the end-to-end optimization of the prompt, allowing task-specific knowledge to be injected into the pretrained model in a lightweight manner. The goal of prefix language modelling is to find a prompt that can effectively guide the model to generate an entire sequence of sentences in response to the given prompt. This approach is particularly useful when generating text for specific tasks or domains.

On the other hand, autoregressive language modelling is a feedforward technique used to train models for sequence-to-sequence natural language processing tasks. In this approach, the complete weights of the input are transferred to the first stage of the decoder. There is no explicit target sentence for the decoder. Instead, the next stage of the decoder is fed with the word it has just output at the previous stage, effectively generating text one token at a time based on the preceding tokens. Autoregressive language modelling is commonly used in tasks such as translation, text summarization, and dialogue generation, where generating fluent and coherent sequences is essential.
