Week 6 Report

2023-12675 박지호

The topic of today's lecture was mistral language model, and the update for this week.

First, we were introduced to a selection of models mistral provides. Premier models included ministral, small, large, pixtral large, codestral, and embed. Free tier models include pixtral, nemo, mamba, mathstral, and some other models. Preimier models cost money to use, so it would be better for students to use free tier models.

Then, the presenter gave us more description about Pixtral, a vision model. Usually, vision models only function on fixed image sizes, however, mistral natively supports arbitrary image sizes. It obviously works faster on smaller images, but it is also accurate for larger models. Also, it is designed to support arbitrary number of images interleaved with texts, unlike traditional models that work only for a single image.

They also gave us some description about the model's structure which allows the model to perform these tasks. The model has layers which handles information in a way that it can perform computation regardless of the image size or count. The Pixtral model's performance metrics were also given, and the model seemingly performed better than other famous models like GPT 4o and llama.

Then, the practical use cases of Pixtral was given. The presenter gave us a live demo where they instructed the Pixtral model figure out the price of a specific item from a receipt, which it was able to successfully. The presenter also made the model to format the data in json format. The presenter also noted some other possible use cases, like captioning the image.

The presenter then moved on to the Codestral model. The codestral model, as its name suggests, focus more on writing codes and such. The presenter then gave us a live demo using a codestral model, where the model generated a html file as instructed. When they asked the model to make changes in the page, the model only edited necessary parts in the html file. The presenter also showed us how they weere able to make python or SQL runner using this feature.

Then, they explained how the model works. The mistal AI uses speculative reasoning to generate fast, but just as accurate output. The model first proposes the speculative tokens using a draft model, and the speculative tokens are passed to a target model as well as original model. The target model's result is then used to verify the speculative tokens. This approach allows the model to get results faster, even if all the speculative tokens are rejected.

This lecture was both highly informative and inspiring, highlighting the advanced capabilities of the Mistral models. The Pixtral model's ability to handle arbitrary image sizes and process multiple images with text represents a significant leap forward in vision models. Its flexibility and superior performance compared to models like GPT-4o and LLaMA demonstrate the impressive advancements achieved by the Mistral team.

The Codestral model's focus on coding tasks was equally compelling. Its ability to generate, modify, and optimize code efficiently provides a glimpse into how AI can streamline software development. The live demonstration of generating and editing HTML, as well as creating Python or SQL runners, showcased its real-world utility and potential to accelerate workflows.

The integration of speculative reasoning to deliver faster and highly accurate results is a particularly innovative feature of the Mistral models. This lecture was an excellent reminder of how rapidly AI technology is advancing and its transformative potential across industries.