WHITE PAPER

Fine-Tuning of Large Language Models

Executive Summary

Using parameter-efficient fine-tuning (PEFT), we enhance Large Language Model (LLM) performance for specific tasks like invoice data extraction. With minimal training data and resources, our adapted model outperforms even GPT-4o – showing a 22% improvement through fine-tuning and performing 16% better than GPT-4o, while being dozens of times smaller – offering a sustainable, customizable, and on-premises-ready AI solution.

INTRODUCTION

Large Language Models (LLMs) are one of the biggest breakthroughs in the world of Artificial Intelligence (AI). These models have impressive generative capabilities and became available and known to the public at first with the release of ChatGPT. LLMs have a broad knowledge and can deal with various tasks due to pretraining with a huge amount of data. But LLMs often underperform on specific use-cases even if the prompts to the model are optimized. Since the training of large models with billions of parameters is very resource-intensive, special methods are needed to increase the performance on use cases and domains.

SETTING

We use parameter-efficient (PEFT) methods to fine-tune pretrained LLMs in a sustainable way. Thus, the weights of the pretrained foundation model are not modified, instead we train a small number of additional parameters (~ 1 % of the original weights) on some hundred use case or domain related training samples. The approach increases the performance of the model on specific use cases or domains without affecting the overall capacity of the model. The fine-tuning allows small LLMs with ~ 7 billion parameters to surpass the performance of bigger models like GPT-4o. Smaller models provide faster responses and are well-suited for on-premises solutions since they require less computational resources – making them not only more cost-efficient but also the more sustainable choice. The reduced resource demand significantly lowers the environmental impact. The additional trained parameters are saved in an adapter which can be dynamically activated. Our adapter for the IDA invoice module is trained on the multi-modal Qwen2.5-VL model which allows the processing of text and images.

ADVANTAGES

- On-premises solutions
- Customization for specific use cases and domains
- Training with few training samples
- Multi-modal models
- Resource-efficient and sustainable

Information extraction from invoices is an important use case of document processing. There are recurrent questions of interest like, e.g., the extraction of the total value. We fine-tuned the foundation model Qwen2.5-VL using the described PEFT methods. The training involved both the extraction of individual entities and the retrieval of invoice tables. Moreover, the LLM was trained to correctly predict the position of the requested entities.

The table below compares the performance of the foundation model with 7 billion parameters (Qwen2.5-VL-7B), the fine-tuned version of this model, and GPT-4o based on the so-called F1 score. Two evaluations are made. On the one hand, the model prediction is marked correct if the predicted text is correct. On the other hand, the prediction is correct if not only the text is correct but also the position matches.

	Text correct	Text & position correct
Qwen2.5-VL-7B	0.72	0.60
Qwen2.5-VL-7B fine-tuned	**0.88**	**0.81**
GPT-4o	0.76	0.66

Table 1: F1-Score on intern invoice QS set

Remarkably, the fine-tuned model not only increases the performance of the foundation model, but even surpasses the performance of GPT-4o, considered to be one of the most powerful LLMs on general tasks. **The fine-tuned model shows a 22% improvement over the base model and performs 16% better than GPT-4o.** These results demonstrate the benefit of LLM fine-tuning for significant improvements on specific use cases with only few data.

ABOUT PLANET AI

PLANET AI is a research-driven company dedicated to developing software products with humaninspired cognitive capabilities for information processing and understanding. By utilizing proprietary deep learning technology, PLANET AI empowers organizations to unlock information trapped in documents, seize digitization opportunities, and eliminate manual effort for data capture. The Intelligent Document Analysis software suite offers comprehensive capabilities for customers with the common desire for short time-to-value automation and high-quality data capture, extraction, and understanding.

PLANET AI serves a variety of customers that include Fortune 500 companies, scanning Service providers, as well as software vendors in business process automation and content management. Since its beginnings in 1992, PLANET AI has established itself as a global technology leader in cognitive computing. In 2023, German IT provider Bechtle acquired a majority share of PLANET AI.