Using GPT-3.5 Turbo to Train Specialized Artificial Intelligence Models

Keywords: Large Language Models (LLMs), Specialzed Artificial Intelligence Models (SAIMs), Prompts, Patents, Claims, Keywords, Technical Features, Augmented Invention, Perception

Author: Kai-Jun Johnson Huang, Taipei Municipal HuaiSheng Junior High School, Grade 9

作者: 黃楷鈞 台北市懷生國中 9 年級

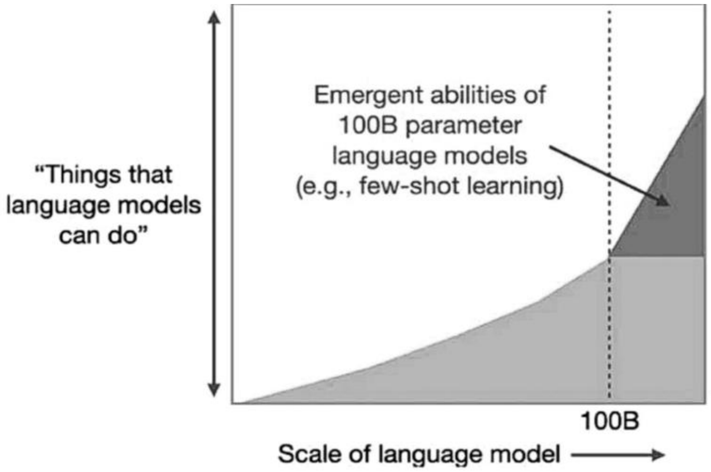
Advisor: Jieh-Sheng Lee, Assistant Professor, National Yang Ming Jiao Tung University

1 Introduction

1.1 Research Background

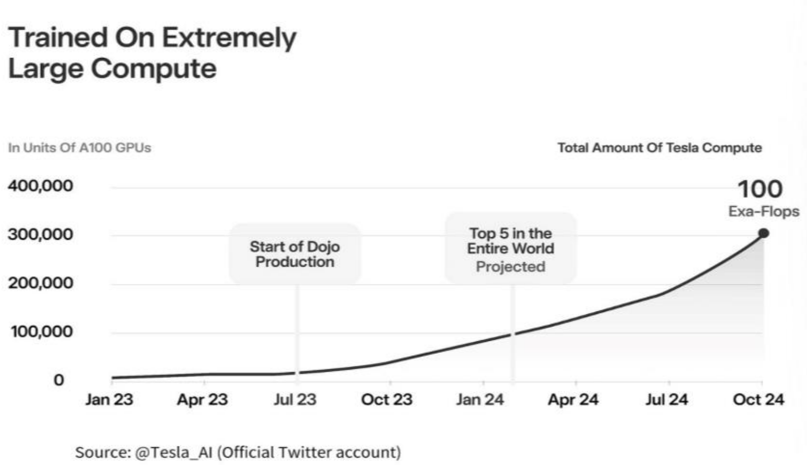
This paper discusses how to train Large Language Models (LLMs) using engineered prompts paired with selected data for creating specialized artificial intelligence models (SAIMs), particularly using keywords to select English patent claims for fine-tuning GPT-3.5 Turbo to generate effective augmented invention SAIMs. I started researching LLMs in early 2022 and presented a paper titled "Generating and Evaluating Patent Claims Using Large Language Models Fine-Tuned with Selected Patents and Patent Applications for Augmented Invention" at the First High School Language Science Fair on October 29, 2022 [1] and demonstrated an operational AIEdison website for augmented invention, winning a Certificate of Merit from Taiwan's National Science and Technology Council. Coincidentally, OpenAI launched ChatGPT based on GPT-3.5 in November 2022, and igniting the LLMs craze, involving Google (PaLM) and Meta (LLaMA). This LLMs development is exciting and promising for linguists since LLMs originated from Natural Language Processing (NLP). However, early attempts using rule-based systems and statistical models limited NLP's scalability and generation capabilities (Jurafsky & Martin, 2009) [2].

Therefore, there was a shift towards representing words as continuous vectors in high-dimensional space, resulting in a breakthrough in NLP. Following this breakthrough, the application of recurrent neural networks such as Long Short-Term Memory networks (Hochreiter & Schmidhuber, 1997) [3] enabled language models to maintain a memory of previous inputs in a sequence. In 2017, Google Brain published a paper titled "Attention Is All You Need" introducing attention mechanisms from a machine language translation project. Attention mechanisms allow models to focus on different parts of the input data to make predictions. Google Brain also introduced the transformer architecture, relying solely on attention mechanisms without recurrence (Vaswani et al., 2017) [4]. With attention and transformer, researchers discovered that given enough data and compute power, language models performed better with increasing model size (Kaplan et al., 2020) [5]. LLMs like OpenAI's GPT, Google's PaLM, and Meta's LLaMA underscored this trend. These models, with hundreds of billions of parameters, can interact with humans, answer questions, and even assist in creative tasks. As we can see from the graph below from a paper titled "Emergent Abilities of Large Language Models" (Wei et al., 2022) [6], for LLMs with more than one hundred billion parameters, their capabilities emerge more rapidly with an increasing number of parameters.


"Things that language models can do"
Emergent abilities of 100B parameter language models (e.g., few-shot learning)
100B
Scale of language model

1.2 Research Motivation

However, LLMs with an exponential increase in parameters may not be ideal, especially for handling specialized tasks. Moreover, the craze of chasing larger LLMs has caused serious concerns. These models demand massive computing resources, raising questions about environmental impact and accessibility for researchers with limited resources. Furthermore, these LLMs can perpetuate biases present in their training data, leading to unfair or inappropriate outputs (Bender et al., 2021) [7]. In fact, GPT-4 launched in March 2023, still hallucinates and invents facts. Thus, it is prudent to search for promising LLM fields that are environmentally and economically friendly to researchers. For example, in July 2023, Tesla started the production of Dojo supercomputer for training its fleet of autonomous vehicles. From the graph provided by Tesla below, it's estimated that Dojo will have compute power equivalent to 300,000 units of A100 GPUs (Hawkins, 2023) [8]. At US$10,000 per A100 GPU, the cost would be US$3 billion if Tesla uses A100 GPUs to build the system for training autonomous vehicles, which is too costly even for Tesla.



Source: @Tesla_AI (Official Twitter account)

In this paper, I discuss an exciting LLM application, which is training LLMs to create specialized artificial intelligence models (SAIMs) capable of handling domain-specific tasks effectively. By fine-tuning GPT-3.5 Turbo, I created augmented invention SAIMs to help innovators invent solutions to make our world better. There are no privacy or copyright concerns since most inventions made in the world have been written as patent applications and have been published in the public domain without copyright. Furthermore, patent applications provide well-defined claim language containing key technical features to manifest the core innovation. According to the World Intellectual Property Organization, each year several million patent applications have been filed worldwide. Thus, we have a humongous volume of patent language documents containing almost all inventions made in modern history, which is ideal for training LLMs. Therefore, I have selected patent data with a specific domain focus to fine-tune GPT-3.5 Turbo for creating effective augmented invention SAIMs, so I can achieve my AI for Good (AI4G) goal through innovation.

2 Relevant Papers

2.1 Papers by Jieh-Sheng Lee

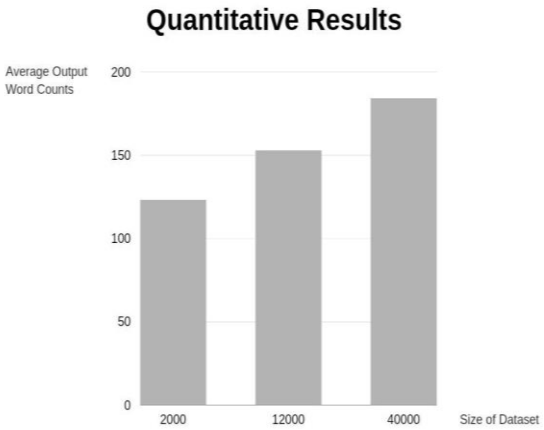
My research advisor, Professor Jieh-Sheng Lee, is one of the pioneering researchers using patent language data to fine-tune pre-trained LLMs. In his paper titled "Patent Claim Generated by Fine-Tuning OpenAI GPT-2," he evaluated the generated patent claims in terms of prediction accuracy (Lee & Hsiang, 2020) [9]. He worked on PatentGPT-J, which was created by pre-training open source GPT-J-6B with USPTO patent data and set out his findings in "Evaluating Generative Patent Language Models," focusing on prediction accuracy analysis again. (Lee, 2022) [10].

Professor Lee's work inspired me to use patent data for fine-tuning LLMs. However, my research differs from Professor Lee's focus on predicting accuracy and concentrates on creating fine-tuned models that can generate innovative patent claims to inspire inventors (augmented invention). To create effective augmented invention models, I have selected patent language data with precise technical focus for training or fine-tuning LLMs since every technical field has its unique requirements and patent documents are very different across different technical fields. Furthermore, my emphasis is on measuring the quality/inventiveness of patent claims generated from the fine-tuned models since my goal is to create effective augmented invention SAIMs, which can produce innovation inspirations in generated claims that are most helpful to innovators.

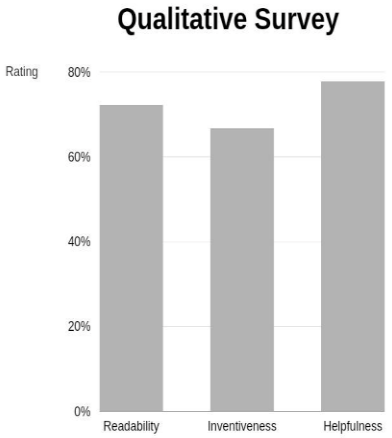
2.2 Author’s Previous Research

In my previous research [1], I selected patent claims with a general vehicle focus and downloaded them via the USPTO public search database [11] and a private patent database, acquiring 40,000 patent claims filed by Toyota with the USPTO from September 2017 to August 2022, to fine-tune GPT-2. From the 40,000 claims, I randomly selected 2,000 claims and 12,000 claims, so in total, I had three claim datasets with 2,000, 12,000 and 40,000 claims. Then, I fine-tuned GPT-2 with different datasets using Hugging Face’s Transformers library and PyTorch at Google Colab Pro+.

I assessed the fine-tuned models and found that the size of datasets corresponded proportionally to the length of generated claims with larger datasets producing longer claims. Furthermore, longer generated claims often contain more technical features. Under the patent laws, technical features are essential elements in determining the inventiveness of patents. Thus, new technical features of the generated claims may inspire inventors looking for innovative solutions. I conducted subjective surveys by asking three experienced patent practitioners to review the generated claims and rate them in terms of inventiveness, helpfulness, and readability. As you can see from the graphs below, quantitative results show that the larger the fine-tune datasets, the longer the generated claims with more technical features and qualitative surveys show that larger fine-tune datasets scored higher on readability, inventiveness and helpfulness consistently. However, this led to an important question of how big the dataset is ideal, and whether it is possible to get good results with smaller datasets.



Larger the fine-tune dataset, longer the generated claims (15/18 - 83.3%), and longer generated claims contain more technical features than shorter claims (83.3%).



Claims generated by GPT-2 fine-tuned with largter datasets score higher on readability (72.2%), inventiveness (66.7%) and helpfulness (77.8%), which was rated by 3 patent lawyers experienced in the field.

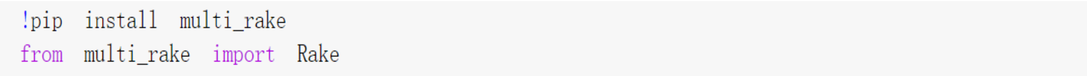
3 Research Method

3.1 Model Choice and Settings

I started researching with GPT-3 but switched to GPT-3.5 Turbo when OpenAI opened it to the public for fine-tuning in August 2023. GPT-3.5 Turbo has 175 billion parameters, which is more than 2 times larger than Meta's LLaMA-2 (70 billion parameters), and it is above the 100 billion parameters emergent capabilities threshold. Although deploying LLaMA-2 locally seems better in terms of privacy and convenience, the deployment cost is much higher than using GPT-3.5 Turbo. Furthermore, all my training datasets are in the public domain without privacy concern. Moreover, GPT-3.5 is a fine-tuned version of GPT-3, with a main feature designed to eliminate toxic outputs by using a 12 stacks of decoder blocks with multi-head attention blocks. Therefore, GPT-3.5 Turbo may be the most suitable LLM for training SAIMs using engineered keyword prompts and scientifically selected patent data. For this research, the LLM settings are same as those used in my previous research, except for Temperature. I experimented with Temperature values 0.1, 0.3, 0.5, 0.7 and 0.9, and I settled at 0.7 to give GPT-3.5 Turbo certain randomness in returning not only the most probable token, so the trained SAIMs could be creative and innovative.

3.2 Data Selection and Prompt Preparation

In comparison with using only general vehicle patent data in the previous research, I use keywords corresponding to important technical features in the vehicle perception field to select relevant patents and obtain 20,000 patents and patent applications published by the USPTO between September 2017 to August 2023. Therefore, I have created two groups of datasets: one group consists of general vehicle patents and the other group consists of vehicle perception specific patents. Using a random selection process, I further create within each group three different sizes of datasets (200, 2,000 and 20,000 patent claims). Then, to prepare prompts pairing with claims in each dataset for training with GPT-3.5 Turbo, I have tried various software tools and eventually use rapid automatic keyword extraction (RAKE) for Python to extract from all patent claims the most relevant technical features serving as keywords for the engineered prompts.



```
!pip install multi_rake
from multi_rake import Rake
```

After completing fine-tuning, a SAIM is generated with each dataset. I pick 5 keywords and use them as the same prompt to input into all SAIMs and obtain 10 outputs per model/SAIM.

4 Research Analysis and Result

4.1 Good Output Rate

Since I always input the same prompt (five vehicle perception related keywords), SAIMs trained with perception specific patents have much higher good output rates than SAIMs trained with general vehicle patents. It is interesting that the keyword "circuitry" has been missing most often from the outputs. Maybe SAIMs have difficulty recognizing and predicting with the keyword.

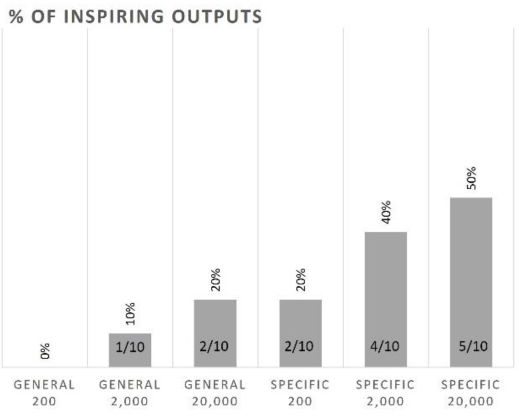
GPT 3.5 Turbo Fine-tuned	Keywords/ Technical Features Input (Prompt) & Output (Claim)					10 Outputs (Generated Claims)	
(Epoch: 5, Temperature: 0.7)	alert	prediction	circuitry	destination	identify	With Defect/ Missing Keyword	% of Good Outputs
Perception Specific Patents	Generated Claims Missing Keywords (times)						
200 Claim Sets	1	1	1			3/10	70%
2,000 Claim Sets			1			1/10	90%
20,000 Claim Sets						0/10	100%
General Vehicle Patents							
200 Claim Sets	1	1	1		1	5/10	50%
2,000 Claim Sets			1		2	4/10	60%
20,000 Claim Sets			1			2/10	80%

SAIMs fine-tuned with perception specific claim datasets have higher good output rates and fewer missing keywords.

4.2 Percentage of Inspiring Outputs

I obtain the same output length result as my previous research and found that the larger the training dataset, the longer the generated claim outputs. However, most excitingly, the percentage of outputs with inspiring new keyword or claim language, SAIMs trained with vehicle perception specific patents significantly outperformed SAIMs trained with general vehicle patents.

GPT 3.5 Turbo Fine-Tuned	Word Count (Generated Claims)	Inspiring New Keyword or Claim Language
(Epoch: 5, Temperature: 0.7)	Average	10 Outputs
Perception Specific Patents		
200 Claim Sets	77	2/10
2,000 Claim Sets	100	4/10
20,000 Claim Sets	115	5/10
General Vehicle Patents		
200 Claim Sets	76	0/10
2,000 Claim Sets	90	1/10
20,000 Claim Sets	112	2/10



5 Research Conclusion and Suggestion

5.1 Conclusion

Through this research, I have found that models fine-tuned with selected vehicle perception patent claim datasets compared to models fine-tuned with general vehicle patent claim datasets can generate a much higher percentage of good claim outputs. Furthermore, these good outputs contain substantially more inspiring new technical features and/or claim language for innovators in the vehicle perception field. In fact, the smallest model trained with 200 selected vehicle perception patent claims (Specific SAIM) performed almost as good as the largest model trained with 20,000 general vehicle patent claims (General SAIM) in terms of good output rate and percentage of inspiring outputs. Moreover, comparing the largest Specific SAIM and the largest General SAIM (both trained with 20,000 claims), the Specific SAIM has 2.5 times more inspiring outputs than the General SAIM, and has 0% bad output rate (no defect/missing keyword) vs. General SAIM's 20%. These findings have proven that training LLMs with domain specific datasets paired with engineered prompts can produce effective SAIMs, and Specific SAIMs can handle domain specific tasks much better than general purpose LLMs and General SAIMs.

5.2 Suggestion

The ChatGPT craze, started in November 2022 may be cooling down. It has been reported that OpenAI suffered a loss of US$540M in 2022 and has been trying to raise additional US$1,000B to develop AGI or even super intelligence. If this is true, I am not optimistic about OpenAI's future. On the other hand, open source LLMs and platforms such as Hugging Face have been flourishing. In fact, Hugging Face on August 23, 2023 raised US$235M from investors, including Google, Nvidia, Amazon, etc. Furthermore, Databricks, a company has its origin in academia and the open source community, raised US$500M on September 13, 2023 (Wilhelm, 2023) [12]. Databricks provides a collaborative cloud platform where developers and researchers can train, manage, and deploy AI applications and models (LLMs). As we can see, open source and collaborative platforms for developers and researchers to carry out AI tasks are prospering. Thus, my next phase of research will move towards training SAIMs with open source LLMs. I have been using Meta's LLaMA 2 and will try to download the 70B pre-trained model to a local machine. I will improve Augmented Invention SAIMs with further fine-tuning at Hugging Face and using the local AI machine.

6 References

[1] Huang, K. J., & Lee, J. S. (2022). Generating and Evaluating Patent Claims Using Large Language Models Fine-Tuned with Selected Patents and Patent Applications for Augmented Invention. Graduate Institute of Linguistics, National Taiwan University. https://taiwan-olympiad-in-linguistics.github.io/scifair/assets/FinalWorks/06-%E4%BD%9C%E5%93%81%2029.pdf

[2] Jurafsky, D., & Martin, J. H. (2009). Speech and Language Processing. Prentice Hall.

[3] Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. Neural Computation.

[4] Vaswani, A. et al. (2017). Attention Is All You Need. Advances in Neural Information Processing Systems (NeurIPS).

[5] Kaplan, J. et al. (2020). Scaling Laws for Neural Language Models. arXiv. Retrieved from https://doi.org/10.48550/arXiv.2001.08361.

[6] Wei, J. et al. (2022). Emergent Abilities of Large Language Models. Transactions on Machine Learning Research. Retrieved from https://openreview.net/pdf?id=yzkSU5zdwD.

[7] Bender, E. M. et al. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT), 610–623. Retrieved from https://doi.org/10.1145/3442188.3445922.

[8] Hawkins, A. J. (2023). Tesla Starts Production of Dojo Supercomputer to Train Driverless Cars. The Verge. Retrieved from https://www.theverge.com/2023/7/19/23800854/tesla-driverless-dojo-supercomputers-production.

[9] Lee, J. S., & Hsiang, J. (2020). Patent Claim Generated by Fine-Tuning OpenAI GPT-2. arXiv. Retrieved from https://doi.org/10.48550/arXiv.1907.02052.

[10] Lee, J. S. (2023). Evaluating Generative Patent Language Models. World Patent Information, 72, 102173. https://doi.org/10.1016/j.wpi.2023.102173.

[11] United States Patent and Trademark Office (n.d.). Patent Public Search. Retrieved from http://www.uspto.gov/patft/index.html.

[12] Wilhelm, A. (2023). Databricks raises $500M more, boosting valuation to $43B despite late-stage gloom. TechCrunch. Retrieved from https://techcrunch.com/2023/09/14/databricks-raises-500m-more-boosting-valuation-to-43b-despite-late-stage-gloom/.