A Survey of Generative Pre-Trained Transformer (GPT) Models

Vaidehi Bulusu
University of California, Berkeley
vaidehi-b@berkeley.edy

INFO 159 Subfield Survey Paper
May 10, 2024

Abstract

In this survey of Generative Pre-Trained Transformer Models, I provide an in-depth exploration of the foundational papers of the field, and an overview of state-of-the-art papers. Much of this work relates to artificial general intelligence, learning process and prompting, and specific applications of GPT.

1 Introduction

The Generative Pre-trained Transformer (GPT) model is an advanced artificial intelligence system designed to understand and generate human-like text based on the input it receives. Developed by OpenAI, GPT belongs to a category of machine learning models known as transformers, which are renowned for their ability to handle sequential data, such as language, with exceptional proficiency.

A distinctive feature of GPT is its training process, which involves two main stages: pre-training and fine-tuning. During pre-training, the model learns the intricacies of language from a vast dataset in an unsupervised manner, absorbing the general structure and patterns of the language without specific instructions. This stage equips GPT with a robust foundational knowledge of language. In the fine-tuning phase, GPT is further trained on a smaller, specialized dataset tailored to specific tasks like answering questions, translating languages, or even generating creative content. This targeted training helps refine the model's responses to be more task-appropriate and accurate.

GPT models have grown significantly in complexity and capability over time, with each version featuring more parameters that allow for deeper understanding and more nuanced language generation. These models are widely used in various applications, from automated customer service chatbots to tools for writing assistance, showcasing their versatility and effectiveness in processing and generating language.

2 Foundations

In this section, I provide an overview of foundational papers about GPT. To produce the GPT as we know it today – specifically, GPT-4, which is the latest one – breakthroughs in unsupervised learning and few-shot learning were required. Radford et al. (2018a) and Radford et al. (2018b) discuss breakthroughs they made in unsupervised learning. To provide context, NLP tasks – from question answering to document classification – require large amounts of data, especially labeled data. While there is an extensive amount of unlabeled text corpora, labeled data is much more limited. This makes it difficult to train discriminative models, while require labeled data. Since GPT is a generative model, unsupervised learning is essential to its success. One solution to the problem is the generative pre-training of language models on a large amount of diverse unlabeled text, followed by fine-tuning on a discriminative model for each task, can significantly improve accuracy (Radford et al., 2018a). This novel approach is different from previous approaches as it uses a task-aware input transformation without significant changes to the model architecture. Hence, this is a task-agnostic model, and it performed better than discriminative models that use task-specific architecture. Specifically, it outperformed the state of the art in 9 out of the 12 tasks studied, with 8.9% improvement on common sense reasoning and 5.7% on question answering. They also found that transformers and data sets containing texts with long-range dependencies work best with this approach.

Radford et al. (2018b) discusses another improvement found with unsupervised learning on task-agnostic datasets, particularly a dataset comprising millions of webpages, called WebText. When the language model is given a document and questions as input, it produces responses that achieve an F1 score of 55 on the CoQA dataset.

This performance is equal to or better than 3 out of 4 baseline systems, even though the model was not trained using the 127,000+ training instances. The language model's capacity is crucial for achieving success in zero-shot task transfer, and increasing it leads to performance improvement in a linear manner across tasks. The GPT-2 model, which has 1.5 billion parameters, is a Transformer that delivers state-of-the-art performance on 7 out of 8 language modeling datasets without any fine-tuning, but it still does not fully capture the complexity of the WebText dataset. The model's samples demonstrate these enhancements and consist of cohesive paragraphs of text. These findings provided a direction for constructing language processing systems that can learn to fulfill tasks based on natural contexts.

While the novel approach discussed Radford et al. (2018a) lead to significant improvements with task-agnostic architecture – which is crucial for the model to be able to solve a wide-variety of problems, like ChatGPT does today – the fine-tuning phase still required a large task-specific dataset with labeled examples. This is in contrast to human learning, which only requires a few examples. Hence, developing the capacity for few-shot learning would make the model exponentially more powerful, allowing it to better approximating human cognition. Brown et al. (2020) address this concern, demonstrating that increasing the size of a language model significantly enhances its ability to perform task-agnostic, few-shot performance, even achieving comparable results to previous state-of-the-art fine-tuning methods. In this study, they train GPT-3, which is an autoregressive language model with 175 billion parameters (10 times more than previous language models) and test its capacity for few-shot learning. The model is applied without any gradient updates or fine-tuning, and the few-shot prompts were provided using text. They find that the model performs well on various datasets, including datasets for translation, question-answering, and tasks that require instantaneous reasoning or adaption to a specific domain (e.g. unscrambling words). Another accomplishment of this model was that it could generate news articles which human evaluators could not easily distinguish from articles written by humans. However, they found that few-shot learning is not as strong on some datasets.

The culmination of these breakthroughs is the powerful GPT-4, introduced in OpenAI et al. (2024). As described in this technical report, this is a multimodal model that can handle input data in both image and text format, and can generate text output. It is a pre-trained Transformer-based model that can accurately predict the next token in a given document. It even exhibited a level of competence comparable to humans based on certain evaluations. For example, it obtained a score that placed it in the highest 10% among test takers on a simulated bar exam. Creating a robust infrastructure and implementing various optimization techniques was crucial to the development of this model. Bubeck et al. (2023) build on this paper, arguing that the GPT-4 model developed by OpenAI is in an entirely different sub-class of GPT models, as it exhibits greatest general artificial intelligence than previous models. To support this claim, they demonstrate the capacity of GPT-4 to solve new and challenging tasks in various domains including language, mathematics, coding, and law, its performance comparable to human performance. Moreover, the model was able to accomplish this without any special prompting. Based on these results, Bubeck et al. argue that GPT-4 can be viewed as an early version of an artificial general intelligence (AGI) system. They also discuss next steps, especially the development of techniques that go beyond next-word prediction.

3 State of the Art

In this section, I discuss the state-of-the-art developments in GPT models. The recent papers fall under 3 main themes: general intelligence, learning process and prompting, and specific applications.

3.1 General Intelligence

Much of the recent work on GPT models has focused on its capacity for general intelligence, that is, its ability to solve a wide variety of reasoning tasks. This essentially refers to the extent to which the model can approximate human intelligence. Qin et al. (2023) investigated the capacity of ChatGPT for general-purpose problem-solving using zero-shot prompting. They evaluated ChatGPT on 20 popular NLP datasets for 7 task categories. They found that ChatGPT performs well on various reasoning tasks (e.g. arithmetic reasoning) but struggles when solving specific tasks such as sequence tagging.

Bang et al. (2023) evaluates ChatGPT using 23 publicly available datasets covering 8 task categories. They specifically evaluated the multitask, multilingual, and multi-modal capacities of ChatGPT and found that it outperforms LLMs with zero-shot prompting, and even fine-tuned models on some tasks. As for its linguistic generalizability, they found that it has limited capacity to generate non-Latin scripts, but is able to understand them. They also found that it is able to generate multimodal content based on text prompts. As for reasoning, they found that ChatGPT achieves an average accuracy of 63.4% in 10 different reasoning tasks, including logical, non-textual, and commonsense reasoning. It also suffers from hallucination. investigates this problem of hallucination in LLMs. These models hallucinate because of a lack of evidential closure, which refers to the model's output not being constrained to evidence-based claims. In the paper, they discuss a framework for constraining LLM output by evidential closure, which involves generating output based on a validated evidence set.

An important aspect of general intelligence is being able to understand language. Qi et al. (2023) discuss this in their paper by investigating the capacity of LLMs to understand converse relations. They create a new benchmark called ConvRe, specific to converse relations, and conduct experiments to determine the capacity of LLMs to match relations and associated text. They found that LLMs rely on shortcut and struggle with their proposed benchmark.

Another paper that explores the ability of LLMs to understand language is Chan et al. (2024). In this paper, they evaluate ChatGPT on its ability to understand inter-sentential relations including temporal, causal, and discourse relations. They perform evaluations on 11 datasets and use tailored prompt templates, including zero-shot and in-context learning templates. They found that ChatGPT has a remarkable ability in understanding and reasoning about causal relations, but does not display similar proficiency in understanding temporal order of events.

While it is capable of identifying the majority of discourse relations with existing explicit discourse connectives, the implicit discourse relation remains a formidable challenge. Concurrently, ChatGPT demonstrates subpar performance in the dialogue discourse parsing task that requires structural understanding in a dialogue before being aware of the discourse relation.

Another aspect of general intelligence is the capacity for moral reasoning. This is the topic of Khandelwal et al. (2024), in which they investigate the moral reasoning capacity of 3 LLMs – ChatGPT, GPT-4, and Llama2Chat-70B – using the Defining Issues Test in various languages. They found that moral reasoning ability for all models is lower for non-European langauges (Hindi and Swahili) compared to European languages (Spanish, Russian, English), and Chinese.

Wachowiak and Gromann (2023) provides another advancement in learning the AGI capacity of GPTs by investigating the ability of GPT-3 to detect metaphorical language and predict the source domain. They use two distinct datasets , and applying various fine-tuning and few-shot prompting approaches. They found that the model attains an accuracy of 65.15% in English and 34.65

3.2 Learning Process and Prompting

In this section, I discuss papers about how GPT models learn, and prompting techniques to help them learn better. An important paper on this topic is Wei et al. (2023) which explores chain-of-thought prompting. This type of prompting involves providing a series of intermeditae reasoning steps. They found that this technique significantly improves the reasoning capacity of LLMs, especially in arithmetic, common sense and symbolic reasoning tasks, even surpassing the results for a fine-tuned GPT-3 with a verifier. Yao et al. (2023a) build on this work by providing a generalization of chain-of-thought prompting, called Tree of Thoughts (ToT). This technique is especially useful for strategic decision-making. It involves providing various paths of reasoning, allowing the LLMs to develop the ability to consider different paths and self-evaluate to choose the best next course of action, as well as the ability for backtracking and look-ahead. They find that this technique significantly improves the model's ability to solve problems on planning or search tasks, achieving a success rate of 74

Liu et al. (2023) is another paper that explores the planning ability of LLMs. They introduce a framework called LLM+P, which takes a natural language description of a planning problem and provides an optimal plan for solving that problem in natural language. They found that this model was able to provide the optimal solution to most problems.

To build on research on prompting, Yao et al. (2023b) explore a new approach, called ReAct, for improving reasoning in LLMs. This approach involves the inducing the model to generate reasoning traces, allowing it to create, track and update action plans, and handle exceptions. It also allows the model to gather knowledge from external sources. They found that this approach is more effective than state-of-the-art baselines, and results in higher human interpretability and trustworthiness over apraohces without reasoning or acting components.

The generalization of this research is a language specifically for prompting LLMs (similar to SQL). This is the idea introduced in Beurer-Kellner et al. (2023). Specifically, they discuss Language Model Programming (LMP) which involves a combination of text prompting and scripting, allowing more precise interaction with LLMs. To enable this, they implement LMQL (Language Model Query Language), which captures various state-of-the-art prompting methods.

Shifting away from prompting, a challenge in implementing LLMs is the requirement of an extensive amount of training data. Hsieh et al. (2023) provide a new approach, called Distilling step-by-step, which requires smaller models and less training data and still outperforms existing larger models. Power et al. (2022) is another paper about the training process of LLMs, making a breakthrough discovery in how they learn. They suggest that neural networks learn through a process called grokking, which improves the model's ability to generalize from patterns, much beyond overfitting.

3.3 Specific Applications

In this section, I discuss the use of GPT models in specific contexts. One exciting application is in simulation as explored by Park et al. (2023). They introduce the idea of generative agents, which are computational entities that simulate human behavior, including cooking, working, creating art and so on. In this realm of applications is the creation of poetry, as discussed by Belouadi and Eger (2023), who successfully trained a GPT model to write English and German poetry.

Along more practical lines, researchers also successfully trained GPT models to generate research output (Boiko et al., 2023), perform relation extraction for financial documents (Rajpoot and Parikh, 2023), and annotate data (Ding et al., 2023).

References

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity.

Jonas Belouadi and Steffen Eger. 2023. Bygpt5: End-to-end style-conditioned poetry generation with token-free language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.

Luca Beurer-Kellner, Marc Fischer, and Martin Vechev. 2023. Prompting is programming: A query language for large language models. *Proceedings of the ACM on Programming Languages*, 7(PLDI):1946–1969.

Daniil A. Boiko, Robert MacKnight, and Gabe Gomes. 2023. Emergent autonomous scientific research capabilities of large language models.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4.

Chunkit Chan, Jiayang Cheng, Weiqi Wang, Yuxin Jiang, Tianqing Fang, Xin Liu, and Yangqiu Song. 2024. Chatgpt evaluation on sentence level relations: A focus on temporal, causal, and discourse relations.

Bosheng Ding, Chengwei Qin, Linlin Liu, Yew Ken Chia, Shafiq Joty, Boyang Li, and Lidong Bing. 2023. Is gpt-3 a good data annotator?

Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alexander Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes.

Aditi Khandelwal, Utkarsh Agarwal, Kumar Tanmay, and Monojit Choudhury. 2024. Do moral judgment and reasoning capability of llms change with language? a study using the multilingual defining issues test.

Bo Liu, Yuqian Jiang, Xiaohan Zhang, Qiang Liu, Shiqi Zhang, Joydeep Biswas, and Peter Stone. 2023. Llm+p: Empowering large language models with optimal planning proficiency.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior.

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. 2022. Grokking: Generalization beyond overfitting on small algorithmic datasets.

Chengwen Qi, Bowen Li, Binyuan Hui, Bailin Wang, Jinyang Li, Jinwang Wu, and Yuanjun Laili. 2023. An investigation of llms' inefficacy in understanding converse relations.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is chatgpt a general-purpose natural language processing task solver?

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018a. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2018b. Language models are unsupervised multitask learners.

Pawan Kumar Rajpoot and Ankur Parikh. 2023. Gpt-finre: In-context learning for financial relation extraction using large language models.

Lennart Wachowiak and Dagmar Gromann. 2023. Does GPT-3 grasp metaphors? identifying metaphor mappings with generative language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1018–1032, Toronto, Canada. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023a. Tree of thoughts: Deliberate problem solving with large language models.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023b. React: Synergizing reasoning and acting in language models.