Large Language Model for automobile

***Fei Ding** [1,*]

APUS Inc.

Abstract

With the introduction of ChatGPT (OpenAI, 2022) from OpenAI, the power of these models to generate human-like text has captured widespread public attention.

The scale of language models has burgeoned, progressing from modest multimillion-parameter architectures like ELMo (Peters et al., 2018) and GPT-1 (Radford et al., 2018), to behemoths boasting billions, even trillions of parameters, exemplified by the monumental GPT-3 (Brown et al., 2020), Switch Transformers (Fedus et al., 2022) , GPT-4 (OpenAI, 2023), PaLM-2 (Anil et al., 2023), and Claude (Claude, 2023) and Vicuna (Chiang et al., 2023).

The expansion in scale has significantly raised hardware requirements, making it exceedingly challenging to deploy models on mobile devices such as smartphones and tablets.

To deploy on cars , we trained a 7-billion-parameter automobile model, which outperforms GPT-3.5 in the automotive domain. Surpassing all models in areas such as automotive maintenance, navigation queries, and beyond.

Keywords: Large Language Model

Contents

∗. *Corresponding author email: `dingfei@email.ncu.edu.cn`

1 Introduction

Evidence suggests that large models exhibit emergent an emergent capability that is absent in smaller models (Wei et al., 2022). A typical example is few-shot prompting. Few-shot prompting significantly expands the range of tasks supported by models and lowers the barrier to entry for users seeking automation for new language tasks. After GPT-3, models grew in size to 280 billion (Gopher, Rae et al., 2021), 540 billion (PaLM, Chowdhery et al., 2022), and 1 trillion parameters (Megatron, Korthikanti et al., 2023). But their attention has been almost exclusively focused on general large language models,GPT-4 (OpenAI, 2023), PaLM-2 (Anil et al., 2023), Claude (Claude, 2023) ,LLaMA (Touvron et al., 2023), Alpaca (Taori et al., 2023), Vicuna (Chiang et al., 2023), and others (Wang et al., 2022; Zhu et al., 2023; Anand et al., 2023). In recent endeavors aimed at training models solely with domain-specific data, the resulting models, albeit significantly smaller, have outperformed general-purpose LLMs in tasks specific to those domains, such as science (Taylor et al., 2022) and medicine (Luo et al., 2022b; Hernandez et al., 2023). These discoveries inspire further advancement in the development of Large Language Model for automobile .

Objective We train a 7 billion-parameter language model tailored to cater to a diverse array of tasks within the automobile sector.

Domain-specific LLMs. The few existing domain-specific LLMs are trained exclusively on domain-specific data sources (Luo et al., 2022a; Taylor et al., 2022), or adapt a very large general purpose model to domain-specific tasks (Singhal et al., 2023; Lewkowycz et al., 2022). Our research aims to train LLMs using a combination of domain-specific and general data. The resulting model performs exceptionally well on domain-specific tasks while also maintaining robust performance on general benchmarks.

Training data. In terms of data processing, we focus on data deduplication and high-quality data..

Architecture The model architecture of Automobile 7B is based on the prevailing Transformer (Vaswani et al., 2017). Nevertheless, we made several modifications .

Tokenizer We use byte-pair encoding (BPE) (Shibata et al., 1999) from SentencePiece (Kudo and Richardson, 2018) to tokenize the data .

Positional Embeddings We use Rotary Positional Embedding (RoPE) (Su et al., 2021) .

Activations and Normalizations We use SwiGLU (Shazeer, 2020) activation function, a switch-activated variant of GLU (Dauphin et al., 2017) which shows improved results.
We implement Layer Normalization (Ba et al., 2016) at the input of the Transformer block, known for its resilience to the warm-up schedule (Xiong et al., 2020). Furthermore, we integrate the RMSNorm technique proposed by (Zhang and Sennrich, 2019), which exclusively computes the variance of input features, enhancing computational efficiency.

Optimizations We use AdamW (Loshchilov and Hutter, 2017) optimizer for training. β_1 and β_2 are set to 0.9 and 0.95, respectively. We use weight decay with 0.1 and clip the grad norm to 0.5. Following a regimen of 2,000 linear scaling steps, the models are primed, gradually ascending to the peak learning rate before transitioning to a cosine decay, tapering down to the minimum learning rate.

To stabilize training and improve the model performance, we normalize the output embeddings.Because we observed that the norms of the heads tend to be unstable. Additionally, the norm of embeddings for rare tokens decreases during training, which disrupts the training dynamics. Moreover, we have found that semantic information is primarily encoded through the cosine similarity of embeddings rather than L2 distance.

Data We added 45% of automotive specialized knowledge books, comprehensive automotive information, automotive repair textbooks, and maintenance manuals for all vehicles to the data used for training the general LLM. First, train the large model with 40% of general data and 10% of automobile data to let it learn language and basic knowledge. Finally, train it with a mixture of 15% general data and 35% automobile data.

Train We use deepspeed zero2. The communication cost of Deepspeed Zero3 is too high, so we do not adopt it. We have conducted performance optimization to enhance GPU computational efficiency and throughput.

2 Results

We juxtapose Automobile 7B against llama and conduct a re-evaluation of all benchmarks using our proprietary evaluation pipeline to ensure impartial comparison. Performance is assessed across a diverse array of tasks categorized as follows:

Popular aggregated results: MMLU (Hendrycks et al., 2020) (5-shot) and BBH (Suzgun et al., 2022) (3-shot) CommonsenseQA (Talmor et al., 2018), ARC-Challenge (Clark et al., 2018)

Code: Humaneval (Chen et al., 2021) (0-shot) and MBPP (Austin et al., 2021) (3-shot)

World Knowledge (5-shot): TriviaQA (Joshi et al., 2017),NaturalQuestions (Kwiatkowski et al., 2019)

Commonsense Reasoning (0-shot): Hellaswag (Zellers et al., 2019), OpenbookQA (Mihaylov et al., 2018), ARC-Easy, PIQA (Bisk et al., 2020),

Reading Comprehension (0-shot): BoolQ (Clark et al., 2019), QuAC (Choi et al., 2018)

Math: GSM8K (Cobbe et al., 2021) (8-shot) with maj@8 and MATH (Hendrycks et al., 2021) (4-shot) with maj@4

Model	MMLU	HellaSwag	WinoG	PIQA	Arc-e	NQ	TriviaQA	HumanEval	MBPP	MATH	GSM8K
LLaMA 2 7B	44.4%	77.1%	69.5%	77.9%	68.7%	24.7%	63.8%	11.6%	26.1%	3.9%	16.0%
LLaMA 2 13B	55.6%	**80.7%**	72.9%	80.8%	75.2%	**29.0%**	**69.6%**	18.9%	35.4%	6.0%	34.3%
Automobile 7B	**56.2%**	**80.5%**	**72.2%**	**80.2%**	**73.3%**	**30.2%**	**69.2%**	**30.3%**	35.2%	**12.2%**	**45.3%**

Table 1: **Comparison of Automobile 7B with llama.** Automobile 7B is comparable to llama2 across all metrics, and excels in the automotive domain compared to ChatGPT.

3 Conclusion

We employed an new training method, starting with a larger proportion of general data before gradually increasing the incorporation of automotive domain knowledge during the

training process. The resulting model performed better than those trained directly on uniformly mixed general and automotive data.
Our model contributes to the ongoing dialog on effective ways to train domain-specific models.
We have presented AUTOMOBILE 7B, a best-in-class LLM for automobile NLP. We've obtained impressive outcomes on general LLM benchmarks, surpassing similar models in automobile tasks. We attribute this to the meticulously curated dataset.We will continue to gather larger-scale automobile domain data to further optimize our model.

References

Yuvanesh Anand, Zach Nussbaum, Brandon Duderstadt, Benjamin Schmidt, and Andriy Mulyar. Gpt4all: Training an assistant-style chatbot with large scale data distillation from gpt-3.5-turbo. *GitHub*, 2023.

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*, 2023.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, et al. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*, 2021.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, 2020.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2023.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. Quac: Question answering in context. *arXiv preprint arXiv:1808.07036*, 2018.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. Boolq: Exploring the surprising difficulty of natural yes/no questions. *arXiv preprint arXiv:1905.10044*, 2019.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Claude. Conversation with Claude AI assistant, 2023.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Yann N Dauphin, Angela Fan, Michael Auli, and David Grangier. Language modeling with gated convolutional networks. In *International conference on machine learning*, pages 933–941. PMLR, 2017.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *The Journal of Machine Learning Research*, 23(1): 5232–5270, 2022.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021.

Evan Hernandez, Diwakar Mahajan, Jonas Wulff, Micah J Smith, Zachary Ziegler, Daniel Nadler, Peter Szolovits, Alistair Johnson, Emily Alsentzer, et al. Do we still need clinical language models? In *Conference on Health, Inference, and Learning*, pages 578–597. PMLR, 2023.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*, 2017.

Vijay Anand Korthikanti, Jared Casper, Sangkug Lym, Lawrence McAfee, Michael Andersch, Mohammad Shoeybi, and Bryan Catanzaro. Reducing activation recomputation in large transformer models. *Proceedings of Machine Learning and Systems*, 5, 2023.

Taku Kudo and John Richardson. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*, 2018.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, et al. Solving quantitative reasoning problems with language models. *Advances in Neural Information Processing Systems*, 35:3843–3857, 2022.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6):bbac409, 09 2022a. ISSN 1477-4054. doi: 10.1093/bib/bbac409. URL `https://doi.org/10.1093/bib/bbac409`.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in bioinformatics*, 23(6):bbac409, 2022b.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. *arXiv preprint arXiv:1809.02789*, 2018.

OpenAI. Introducing chatgpt. *Blog post openai.com/blog/chatgpt*, 2022.

OpenAI. Gpt-4 technical report. *ArXiv*, abs/2303.08774, 2023.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. corr abs/1802.05365 (2018). *arXiv preprint arXiv:1802.05365*, 2018.

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

Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*, 2021.

Noam Shazeer. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020.

Yusuxke Shibata, Takuya Kida, Shuichi Fukamachi, Masayuki Takeda, Ayumi Shinohara, Takeshi Shinohara, and Setsuo Arikawa. Byte pair encoding: A text compression scheme that accelerates pattern matching. 1999.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *arXiv preprint arXiv:2104.09864*, 2021.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, , and Jason Wei. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*, 2022.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *arXiv preprint arXiv:1811.00937*, 2018.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. Alpaca: A strong, replicable instruction-following model. *Stanford Center for Research on Foundation Models. https://crfm. stanford. edu/2023/03/13/alpaca. html*, 3(6):7, 2023.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*, 2022.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aur'elien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *ArXiv*, abs/2302.13971, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008, 2017.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language model with self generated instructions. *arXiv preprint arXiv:2212.10560*, 2022.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.

Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tieyan Liu. On layer normalization in the transformer architecture. In *International Conference on Machine Learning*, pages 10524–10533. PMLR, 2020.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*, 2019.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. *Advances in Neural Information Processing Systems*, 32, 2019.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.