RESEARCH INTEREST

Chenmien Tan
University of Edinburgh
chenmien.tan@ed.ac.uk

Trustworthy Language Models Large language models have exhibited the ability to acquire massive knowledge from the pre-training corpora and demonstrated promising performance in knowledge-intensive tasks. However, there is no guarantee for the factuality of the generation, which prevents the application of language models in risk-sensitive areas such as education and healthcare. A stright-forward method to maintain the trustworthiness is to fine-tune the chat model on the facts where it makes errors. However, for application usage, chat models are usually instruction-tuned and aligned with human values while the standard fine-tuning can hurt the instruction following abilities and the harmlessness of the model (Gu et al., 2024). There are two potential ways to tackle with the problem: (i) Applying multi-objective learning, *e.g.*, meta learning (De Cao et al., 2021; Tan et al., 2024), to correct the model errors with minimal performance degradation. An inspiration of this approach is that the performance regression on certain tasks led by alignment can be greatly reduced by mixing the alignment and pre-training updates (Ouyang et al., 2022); (ii) Augmenting the generation via retrieving passages from external documents. One downside of the approach is resource-intensive, as the passages are usually much longer than the prompt and generation. The passages also often involve much content irrelevant to the prompt, possibly distracting the model (Yoran et al., 2024). Inspired by LM-supervised retrieval (Shi et al., 2023), a small-scale sequence-to-sequence model can be trained using feedback from large language models to extract useful information from passages, reducing the cost of retrieval augmented generation.

Alignment and Weak-to-Strong Generalization Alignment has become a standard step to build chat language models while it remains ambiguous how to align efficiently. The state-of-the-art language models are aligned through an iterative style (Ouyang et al., 2022; Touvron et al., 2023): since the reward model is typically trained using the generation from the last chat model checkpoint, the quality of the predicted reward will decrease along with the chat model is trained to produce more preferred generation. However, it is unclear when to collect human preference to update the reward model. Drawing inspiration from out-of-domain detection, a confidence score can be designed so that the reward model can actively query human labeling in an online fashion (Muldrew et al., 2024) to minimize disturbance to humans. Furthermore, a decoding mechanism may be designed to selectively produce generations that requires labeling, avoiding unnecessary computation. Alignment also inspires another interesting topic, *i.e.*, weak-to-strong generalization. For stability consideration, the scale of the reward model is typically smaller than the generative model (Ouyang et al., 2022), where supervision signals from a weaker model are used to train a stronger model. It is intriguing to consider whether such paradigm can be extended to broader tasks as language models have exhibited human-level performance on many tasks so that it is difficult to further learn from human annotations (Touvron et al., 2023; Burns et al., 2023). A potential approach is training discriminate models from human preference, enabling them to supervise the generative models to interact with the environment.

Data Influence and Selection The success of instruction tuning reveals that capabilities can generalize well across tasks, but the understanding for the generalization is primarily qualitative rather than quantitative. For example, a common view is that training on code improves the reasoning ability of language models because models trained on code generally exhibit better performance on reasoning tasks, which is akin to observation based on re-training. If there is a tractable proxy that describes the impact of including or excluding data in re-training, it can greatly aid in data selection (Anand et al., 2023; Xia et al., 2024). Unfortunately, the effectiveness of existing proxies is unsatisfying (Schioppa et al., 2023), and they are typically gradient-based and have the same dimension with the model parameters, making it expensive to compute and storage for large language models. The current main-stream method for dimensionality reduction is random projection (Park et al., 2023; Xia et al., 2024), which does not rely on the distribution of the gradient representations

and is ambiguous in performance loss. Inspired by the success of low-rank adaptation (Hu et al., 2022), the gradients on warmed low-rank adapters and the representations after further unsupervised dimensionality reduction may raise as competitive alternatives for the randomly projected gradients.

REFERENCES

Nikhil Anand, Joshua Tan, and Maria Minakova. Influence scores at scale for efficient language data sampling. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. URL https://aclanthology.org/2023.emnlp-main.152.

Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, Ilya Sutskever, and Jeff Wu. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision, 2023. URL https://cdn.openai.com/papers/weak-to-strong-generalization.pdf.

Nicola De Cao, Wilker Aziz, and Ivan Titov. Editing factual knowledge in language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021. URL https://aclanthology.org/2021.emnlp-main.522.

Jia-Chen Gu, Hao-Xiang Xu, Jun-Yu Ma, Pan Lu, Zhen-Hua Ling, Kai-Wei Chang, and Nanyun Peng. Model editing can hurt general abilities of large language models. *arXiv preprint arXiv:2401.04700*, 2024. URL https://arxiv.org/abs/2401.04700.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

William Muldrew, Peter Hayes, Mingtian Zhang, and David Barber. Active preference learning for large language models. *arXiv preprint arXiv:2402.08114*, 2024. URL https://arxiv.org/abs/2402.08114.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html.

Sung Min Park, Kristian Georgiev, Andrew Ilyas, Guillaume Leclerc, and Aleksander Madry. Trak: Attributing model behavior at scale. 2023. URL https://openreview.net/forum?id=PBRArApxMh.

Andrea Schioppa, Katja Filippova, Ivan Titov, and Polina Zablotskaia. Theoretical and practical perspectives on what influence functions do. In *Advances in Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=gGl0n7Onug.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. Replug: Retrieval-augmented black-box language models. *arXiv preprint arXiv:2301.12652*, 2023. URL https://arxiv.org/abs/2301.12652.

Chenmien Tan, Ge Zhang, and Jie Fu. Massive editing for large language models via meta learning. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=L6L1CJQ2PE.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee,

Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, RobertStojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. URL https://arxiv.org/abs/2307.09288.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. Less: Selecting influential data for targeted instruction tuning. *arXiv preprint arXiv:2402.04333*, 2024. URL https://arxiv.org/abs/2402.04333.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. Making retrieval-augmented language models robust to irrelevant context. *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=ZS4m74kZpH.