Boyuan Chen

boyuan.chen.byc@gmail.com Google Scholar https://cby-pku.github.io/ cby-pku

Research Statement

My research is driven by the goal of constructing safe and trustworthy AI systems and achieving scalable oversight & weak-to-strong alignment. Specifically, I am deeply interested in the *hard problem of alignment*:

How to align systems smarter than humans and how to align them on tasks challenging for human evaluation?

The *hard alignment problem* has been extensively studied, with **Scalable Oversight** being a notable method [3, 10, 6, 12]. This approach uses AI to help humans supervise large-scale models. However, scalable oversight largely depends on the principle that evaluation is easier than generation, a concept that may not always hold. Furthermore, the current scalable oversight framework has its limitations. For instance, the Iterated Distillation and Amplification (IDA) [3] and Recursively Reward Modeling (RRM) [10] methods heavily rely on task decomposition, making it difficult to find the optimal policy. Furthermore, RRM is prone to reward over-optimization and reward hacking, complicating the iterative improvement of these frameworks. The debate framework [6, 1] heavily relies on the principle of honesty, which is challenging to implement in practical settings with limited time and contextual windows.
My research focuses on addressing the *hard alignment problem* through a weak-to-strong alignment approach, aiming to employ weaker or smaller models to supervise stronger models. In my view, weak-to-strong alignment encompasses three meanings:

- Enhancing the capability of weak models to match the supervised strong models, thereby achieving supervision. This is also an idea behind scalable oversight.
- Directly using weak supervision signals from weak models to fine-tune strong models, which involves two scenarios:
  - Using mislabels provided by weak models for strong models to learn from in the same type of tasks, as in OpenAI's recent *Weak-to-Strong Generalization* [2]. The trade-off here is whether the strong model merely mimics the weak model, leading to a decline in performance, or leverages its reasoning abilities to learn something truly useful and thereby improve.
  - The second scenario involves using ground truths provided by weak models in tasks of different types or varying difficulty levels (*i.e.*, where a weak model can perform well in task A but cannot complete task B, which the strong model needs to accomplish. The question is whether the labels from A can be generalized to B) akin to humans applying knowledge learned from one context to another in a form of analogical reasoning.
- The third approach to weak-to-strong alignment involves using weak models as functional add-ons to assist in the learning of strong models.

To verify the effectiveness of weak-to-strong alignment, my research initially starts from the safety alignment setting. I first participated in the BeaverTails project, which aims to separate annotations of **helpfulness** and **harmlessness** for question-answering pairs, thus offering distinct perspectives on these crucial attributes [8]. We further showcase applications of BeaverTails in content moderation and RLHF, emphasizing its potential for practical safety measures in LLMs. Subsequently, I engaged in the SafeRLHF project, which aims to introduce the cost model into large language model scenarios, balancing the trade-off between helpfulness and harmlessness [4]. Through these two projects, I have developed a systematic understanding and methodology for the Safety Alignment of LLMs.

To provide a comprehensive and up-to-date overview of the alignment field, my work delves into the core concepts, methodologies, and practices of alignment [9]. Specifically, we identify four principles as the key objectives of AI alignment: **Robustness, Interpretability, Controllability, and Ethicality (RICE)**. Guided by these four principles, we outline the current landscape of alignment research and divide it into two key components: **forward alignment** and **backward alignment**. On forward alignment, we discuss techniques for learning from feedback and learning under the distribution shift. On backward alignment, we discuss assurance techniques and governance practices. The process of conducting the alignment survey has broadened my perspective on AI alignment and further fueled my passion for research.

My recent research focus has been on exploring and validating the path for weak-to-strong alignment. To address the challenge of the *hard alignment problem*, my work introduces a novel method: *Aligner*, which uses a weak model as a correction module [7]. This approach is inspired by the principle that evaluation/correction is sometimes easier than generation. *Aligner* corrects based on the responses from a preceding model and serves as a supervisory signal for training stronger models. As Isaac Newton said, *If I have seen further, it is by standing on the shoulders of giants.* Meanwhile, *Aligner* functions as a model-agnostic plug-and-play module, allowing for its direct application on different open-source and API-based models. **Once trained, the *Aligner* can be applied across different upstream LLMs without requiring parameter adjustments.** Experiments showed that the *Aligner*-7B model enhances both the helpfulness and harmlessness across a spectrum of 11 models, including API-based, open-source, and safety-aligned/safety-unaligned models. Experiment results demonstrate that the *Aligner*-7B increased GPT-4's helpfulness by 17.5% and its harmlessness by 26.9%.

Aligner has received widespread attention from the community; here are some additional pieces of community evidence: Existing work is being conducted around *Aligner*'s unique learning paradigm [13]. The technology company Align-Inc [1] used a smaller version of *Aligner*-2B following Claude3 Opus and achieved the second place on the Alpaca-Eval leaderboard, only behind GPT-4 Preview. This has sparked extensive discussion within the community.

In the future, I will continue to focus on the *hard alignment problem* and use weak-to-strong alignment to tackle these issues. Specifically, I aim to **uncover the mechanisms** of weak labels when fine-tuning the strong model. I will also **explore more effective methods** for weak-to-strong alignment. Moreover, as the saying goes, *all roads lead to Rome*; there is more than one way to solve the *hard alignment problem*. Approaches like Cooperative Inverse Reinforcement Learning (CIRL) [5] still offer valuable solutions, which are also among my research directions. Solving the *hard alignment problem* will be a milestone in tackling the superalignment issue [11], and I will dedicate myself to this endeavor. Despite being a sophomore undergraduate with potentially less insight into the alignment problem, I will leverage my youth and curiosity to seize more opportunities and time for an in-depth exploration of alignment.

References

Brown-Cohen, J., Irving, G., & Piliouras, G. (2023). Scalable ai safety via doubly-efficient debate. *arXiv preprint arXiv:2311.14125*.

Burns, C., Izmailov, P., Kirchner, J. H., Baker, B., Gao, L., Aschenbrenner, L., ... others (2023). Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. *arXiv preprint arXiv:2312.09390*.

Christiano, P., Shlegeris, B., & Amodei, D. (2018). Supervising strong learners by amplifying weak experts. *arXiv preprint arXiv:1810.08575*.

[1] `https://www.alignllm.com/`

Dai, J., Pan, X., Sun, R., Ji, J., Xu, X., Liu, M., ... Yang, Y. (2024). Safe RLHF: Safe reinforcement learning from human feedback. In *The twelfth international conference on learning representations.* Retrieved from https://openreview.net/forum?id=TyFrPOKYXw

Hadfield-Menell, D., Russell, S. J., Abbeel, P., & Dragan, A. (2016). Cooperative inverse reinforcement learning. *Advances in neural information processing systems*, *29*.

Irving, G., Christiano, P., & Amodei, D. (2018). Ai safety via debate. *arXiv preprint arXiv:1805.00899*.

Ji, J., Chen, B., Lou, H., Hong, D., Zhang, B., Pan, X., ... Yang, Y. (2024). Aligner: Achieving efficient alignment through weak-to-strong correction. *arXiv preprint arXiv:2402.02416*.

Ji, J., Liu, M., Dai, J., Pan, X., Zhang, C., Bian, C., ... Yang, Y. (2023). Beavertails: Towards improved safety alignment of LLM via a human-preference dataset. In *Thirty-seventh conference on neural information processing systems datasets and benchmarks track.* Retrieved from https://openreview.net/forum?id=g0QovXbFw3

Ji, J., Qiu, T., Chen, B., Zhang, B., Lou, H., Wang, K., ... others (2023). Ai alignment: A comprehensive survey. *arXiv preprint arXiv:2310.19852*.

Leike, J., Krueger, D., Everitt, T., Martic, M., Maini, V., & Legg, S. (2018). Scalable agent alignment via reward modeling: a research direction. *arXiv preprint arXiv:1811.07871*.

Leike, J., & Sutskever, I. (2023). Introducing superalignment. *OpenAI Blog*.

Saunders, W., Yeh, C., Wu, J., Bills, S., Ouyang, L., Ward, J., & Leike, J. (2022). Self-critiquing models for assisting human evaluators. *arXiv preprint arXiv:2206.05802*.

Yang, K., Liu, Z., Xie, Q., Zhang, T., Song, N., Huang, J., ... Ananiadou, S. (2024). Metaaligner: Conditional weak-to-strong correction for generalizable multi-objective alignment of language models. *arXiv preprint arXiv:2403.17141*.