# The Power of Weak Supervision: An Exploration of Noise Structures, Evolution, and Realism

Yu Meng, Assistant Professor
yumeng5@virginia.edu
Department of Computer Science, University of Virginia

## Overview

In recent years, AI models have showcased remarkable performance and versatility across a wide range of tasks. With the rapid progress in this field, the emergence of superintelligence in the near future seems increasingly likely. However, training such superhuman systems poses significant challenges, particularly due to the absence of explicit superhuman supervision. In this proposal, we focus on exploring key questions in superalignment:

1. What are the effective learning strategies (*e.g.*, linear probes, noise-robust training objectives) to elicit salient concepts from strong student models given the noise structures in weak supervision?

2. With the progression of human knowledge, potentially accelerated by the emergence of superintelligence, weak supervision is likely to evolve. How can we model this evolution to better align student models?

3. How does synthetic weak supervision compare to human supervision, and can synthetic supervision, potentially generated by future superhuman models, be leveraged to enhance the training of student models?

Answering these questions will provide valuable insights into the effective training of superintelligent models. In the subsequent sections, we outline specific tasks to address these questions and propose experimental setups and potential solutions. Collectively, these tasks aim to contribute to the advancement of training methodologies for superintelligent systems.

## Task I: Learning Strategies for Salient Concepts with Weak Supervision

One of the intriguing findings in [Burns et al., 2023, Kirichenko et al., 2023] is that fine-tuning models on weak labels results in linearized representations with respect to ground-truth labels. This suggests that the model learns salient features that are more aligned with the desired concepts, despite the noise in the labels. While this highlights the potential of eliciting salient concepts from strong models through weak supervision, its applicability to aligning future superhuman models raises a critical question:

*Do linear ground-truth concept representations arise from a sufficiently strong student model, specific noise structures in weak supervision, or a combination of both?*

If a strong student model is pivotal, then future larger models will likely learn linearized superhuman concept representations well under human supervision, simplifying the process to be learning a linear probe on top of it. Conversely, if noise structure plays a crucial role, understanding how noise patterns vary across tasks and how model-generated label noise resembles human-generated label noise becomes essential. In this scenario, linear probes alone may be insufficient, necessitating fine-tuning with specially designed objective functions. An overview of our proposed method is illustrated in Figure 1.

**Preliminary Investigation.** Our initial step is to ascertain the primary factors influencing the linearity of ground-truth concept representations. We will train models of various sizes across various tasks (*e.g.*, NLP, chess puzzles, reward modeling) and assess the linearity of their learned representations with respect to ground-truth labels. Determining whether linearity correlates more with model sizes or noise structures across tasks will be crucial for developing new methods to elicit salient concepts with weak supervision.

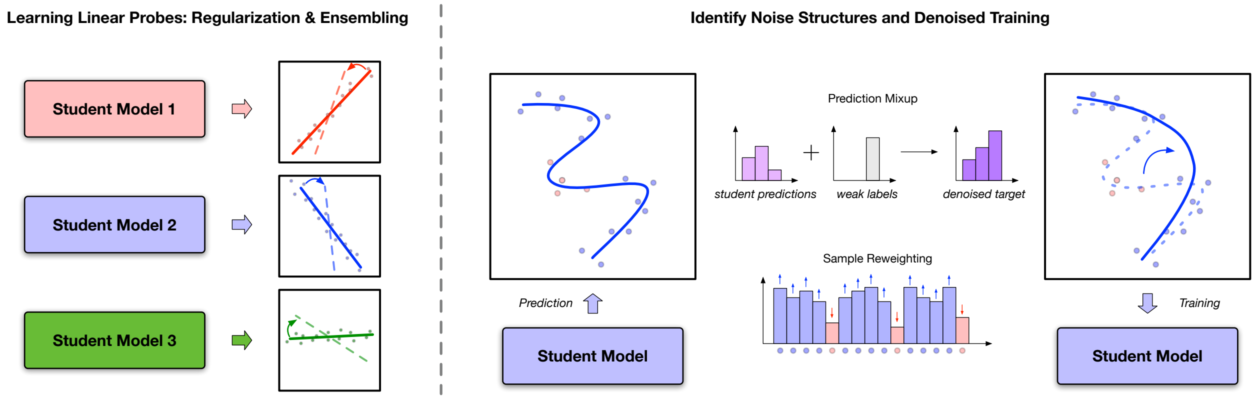

**Figure 1:** (Left) Learning linear probes by regularizing and ensembling the representations from multiple student models. (Right) Identifying noise structures (denoted as red dots) in different settings and new training methods for noise-robustness.

**Proposed Method 1: Learning Linear Probes Under Weak Supervision.** Even if the strong student model produces linearized representations in superhuman concepts, it is still nontrivial to learn the linear probe in the absence of ground-truth superhuman labels. However, exploiting the linearity of learned representations offers opportunities for *regularization* and *ensemble methods*. For example, suppose we have multiple student models or snapshots of model parameters at different training steps, each potentially learning different representations but all expected to be linear with respect to the ground-truth concepts. In that case, regularization techniques can be employed to encourage the linear relations among these representations during training. Additionally, ensembling predictions from multiple noisy linear probes might be beneficial for denoising and training a new and better linear probe. By combining predictions from different linear probes, we can reduce the impact of noise in individual probes, potentially leading to more accurate and robust concept representations [Laine and Aila, 2017, Meng et al., 2022, Nguyen et al., 2020].

**Proposed Method 2: Noise Structures in Weak Supervision and New Training Objectives.** Given the considerable variation in weak-to-strong generalization results across diverse tasks (*e.g.*, NLP, chess puzzles, reward modeling) [Burns et al., 2023], it is very likely that weak supervision exhibits distinct noise patterns for different tasks, impacting the model's learning dynamics differently. It would be beneficial to pinpoint and characterize how these noise patterns arise from various factors, such as model capacity and pretraining data. Specifically, noise in weak supervision may stem from *limited model capacity* [Wei et al., 2022], or *inadequate knowledge acquisition* during pretraining to perform tasks accurately. These scenarios parallel the sources of noise in human supervision, such as cognitive limitations (*e.g.*, supervision from third graders) or lack of domain expertise (*e.g.*, supervision from non-experts for technical tasks). Therefore, understanding and mitigating these noise structures through specially designed training objectives could potentially aid in developing superhuman models with human supervision. Promising solutions likely involve utilizing the strong student model's capacity to denoise the weak labels, akin to the motivation behind auxiliary confidence loss [Burns et al., 2023]. For example, we might use mixed-up predictions from the student model and the weak supervision as the training signal, or train the student model to explicitly identify and downweight noisy labels [Ren et al., 2018, Shu et al., 2019].

**Expected Outcomes.** At the end of this task, we expect the following outcomes: (1) identifying the key factor responsible for the emergence of linearized ground-truth concept representations under weak supervision; (2) designing robust methods (*e.g.*, regularization, ensembling) to learn linear probes over salient concept representations without ground-truth labels; and (3) characterizing various noise structures in weak supervision across diverse scenarios and developing new training objectives to improve noise-robustness during model training.

## Task II: Evolutionary Weak Supervision: Anticipating the Future

With the advent of superhuman models, the pace of human and world knowledge evolution will accelerate significantly. Initially, human beings will leverage these superhuman models to enhance their knowledge, akin to the objective of scalable oversight methods [Cotra, 2021]. Subsequently, successive generations of superhuman models are expected to be trained with the assistance of their predecessors instead of solely from human supervision. This evolution suggests that weak supervision will also progress over time, potentially becoming more robust and effective. Consequently, a critical area for exploration is effective weak-to-strong generalization under evolving weak supervision paradigms. An overview of our proposed method is illustrated in Figure 2.

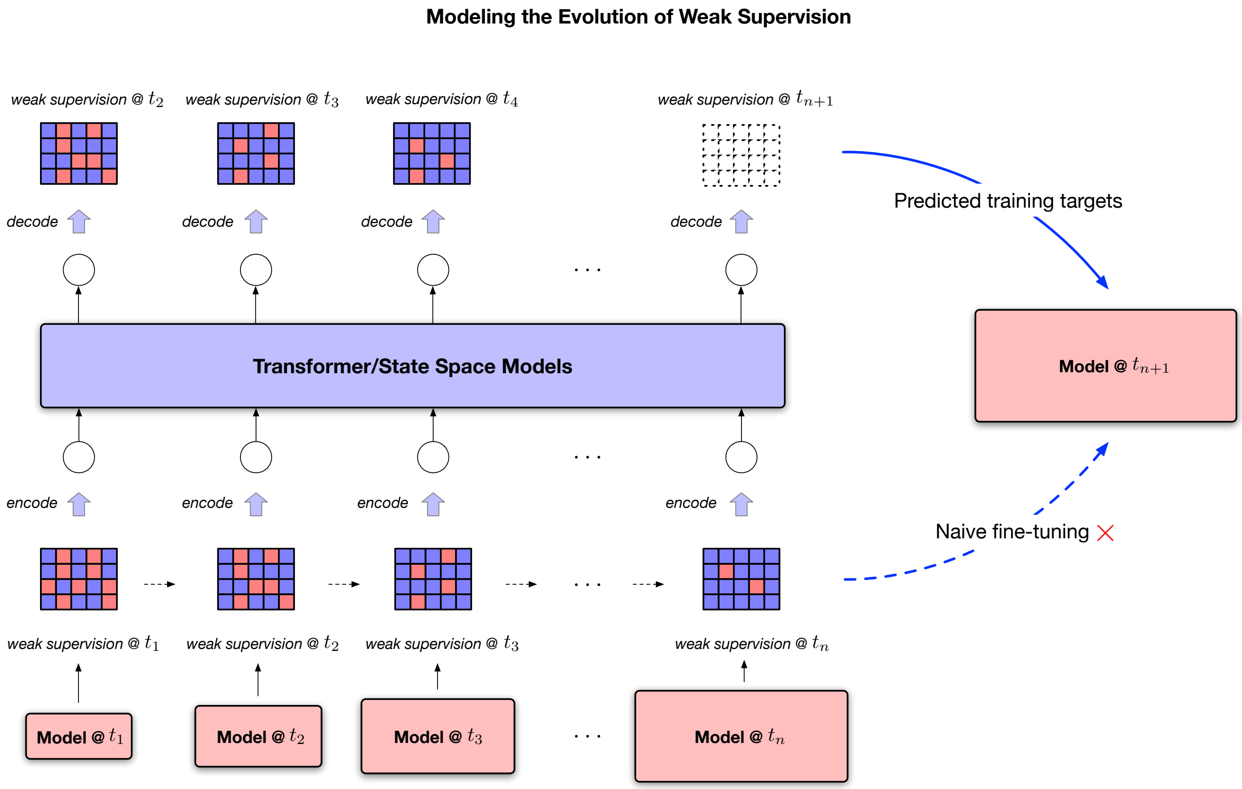

**Figure 2:** Human knowledge progresses over time, and we can simulate this evolution by generating predictions using a sequence of models with varying capabilities. By explicitly modeling the evolution of weak supervision, we can anticipate future weak supervision, which could serve as a more effective training target than naive fine-tuning.

**Experiment Setup.** To simulate the evolution of weak supervision, we propose the following approaches: (1) Training a series of models of varying sizes and utilizing them to generate weak labels. We assume that labels produced by larger models will be more accurate, mirroring the "bootstrapping with intermediate model sizes" setup in [Burns et al., 2023]; (2) Training multiple instances of a fixed-size model on corpora with different knowledge depths (*e.g.*, up to elementary school, up to high school, up to college) and using them to generate weak labels. We assume that labels generated by models with broader knowledge coverage will be more accurate. Our goal is to develop new methods that are more effective than naive bootstrapping (*i.e.*, directly fine-tuning each model on the weak labels generated by the predecessor model in the sequence [Burns et al., 2023]).

**Proposed Method.** The expansion of human knowledge frontiers is typically propelled by more than just comprehending existing knowledge—it also involves gaining insights from the *dynamics of scientific progress* and the *factors influencing the evolution of knowledge*. Therefore, instead of solely learning from the weak supervision itself, we aim to understand the dynamics of weak supervision evolution. One approach is to treat the evolution of weak supervision as a *sequence modeling problem*, where weak supervision at each time step is assumed to be generated based on the previous ones. Under this setup, predicting future weak supervision could be a more effective training target than using the latest available weak supervision alone. We can featurize weak supervision at each time step and use Transformer models or state space models [Gu and Dao, 2023] as the backbone to learn the transition patterns over time. Another possible direction is to apply *meta-learning* to train models to adapt to varying levels of weak supervision quality and knowledge coverage. The goal is to enable the model to quickly learn from emerging weak supervision and achieve more robust and effective weak-to-strong generalization with evolving weak supervision.

**Expected Outcomes.** At the end of this task, we expect the following outcomes: (1) establishing the experiment protocols to simulate the evolution of weak supervision; (2) applying sequence modeling architectures and training methods to predict future weak supervision as better training targets; and (3) developing meta-learning frameworks to train models that can efficiently and effectively adapt to emerging weak supervision.

## Task III: Exploring the Differences: Synthetic vs. Human Weak Supervision

In the above scenarios studied for weak-to-strong generalization, the weak supervision typically consists of model-generated synthetic signals. To ensure the applicability of our findings and proposed methods in building real superhuman models, we must either (1) validate that synthetic weak supervision can effectively mimic human supervision, or (2) demonstrate that synthetic supervision can perform at least as effectively as human supervision in training future superhuman models. This task explores possibilities on both fronts. An overview of our proposed methods is illustrated in Figure 3.

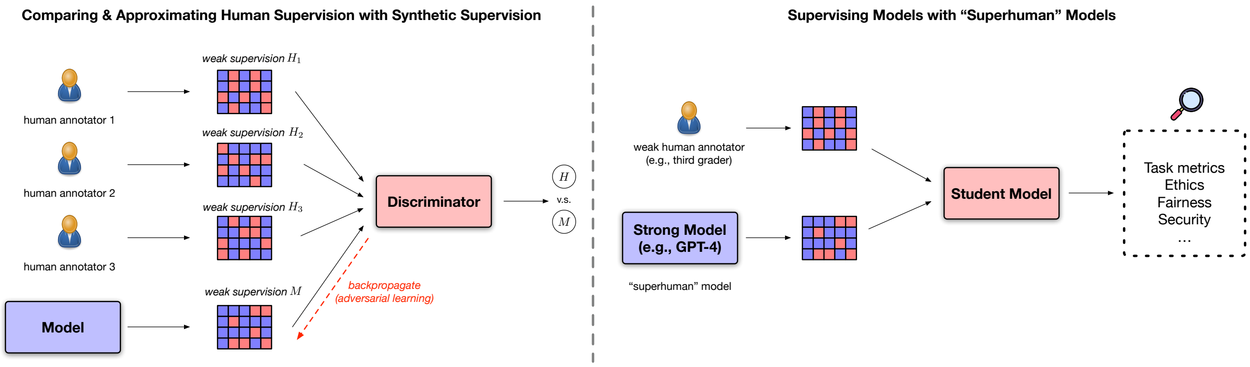

**Figure 3:** (Left) Comparing human supervision with synthetic supervision with a discriminator. The model generating weak supervision might be updated with adversarial learning to enhance its resemblance to human supervision. (Right) We can simulate supervising future models with superhuman models by obtaining human supervision from weak annotators and employing a strong model to generate synthetic supervision, which helps us understand the implications of using superhuman supervision in model training, from the aspects of task metric, ethics, fairness, security and so on.

**Proposed Method 1: Comparing and Approximating Human Supervision with Synthetic Supervision.** Our primary objective is to investigate the feasibility of synthetic supervision closely resembling human supervision across diverse tasks (*e.g.*, NLP, chess puzzles, reward modeling), and identify the key factors necessary to achieve this resemblance. However, quantifying the similarity between synthetic and human supervision poses challenges, particularly in "fuzzy" tasks with subjective or open-ended labels, where even human annotators may disagree significantly (*e.g.*, reward modeling). To this end, we propose training a discriminator model to distinguish between human and synthetic supervision. A lower accuracy of the discriminator indicates more realistic synthetic supervision. We plan to explore various factors influencing the discriminator's accuracy, such as model sizes, pretraining data, and task-specific fine-tuning on human data. Additionally, we may use *adversarial learning* to backpropagate signals from the discriminator to update the model, further enhancing the resemblance of synthetic supervision to human supervision.

**Proposed Method 2: Supervising Models with "Superhuman" Models.** When superhuman models become more prevalent, they are expected to provide supervision of higher quality than what humans can offer. This development raises the question of whether these models will play a dominant role in supervising future generations of models, alongside human supervisors. If so, it will be less concerning if such superhuman supervision does not sufficiently resemble human supervision. Rather, it will be more important to understand the outcomes of training new models with superhuman supervision. To study this, we propose a method where we obtain supervision from weak human labelers (*e.g.*, third graders) and treat that as the standard human supervision. We then use a strong enough model, such as GPT-4 [Achiam et al., 2023], and treat that as a "superhuman" model. With this setup, we train new student models under both weak human supervision and "superhuman" supervision to compare their performance and understand the implications of using superhuman supervision in model training. In addition to task metrics, ethical, fairness, and security concerns arising from training with "superhuman" supervision may need consideration, potentially requiring regulation through learning from human preference [Ouyang et al., 2022].

**Expected Outcomes.** At the end of this task, we expect the following outcomes: (1) validating whether synthetic supervision can sufficiently resemble human supervision and what factors influence such resemblance; (2) developing adversarial training methods to enhance the resemblance of synthetic supervision to human supervision; and (3) studying the outcomes of training models with "superhuman" supervision.

# References

[Achiam et al., 2023] Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. (2023). GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

[Burns et al., 2023] Burns, C., Izmailov, P., Kirchner, J. H., Baker, B., Gao, L., Aschenbrenner, L., Chen, Y., Ecoffet, A., Joglekar, M., Leike, J., et al. (2023). Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. *arXiv preprint arXiv:2312.09390*.

[Cotra, 2021] Cotra, A. (2021). The case for aligning narrowly superhuman models. In *AI Alignment Forum*.

[Gu and Dao, 2023] Gu, A. and Dao, T. (2023). Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*.

[Kirichenko et al., 2023] Kirichenko, P., Izmailov, P., and Wilson, A. G. (2023). Last layer re-training is sufficient for robustness to spurious correlations. In *International Conference on Learning Representations*.

[Laine and Aila, 2017] Laine, S. and Aila, T. (2017). Temporal ensembling for semi-supervised learning. In *International Conference on Learning Representations*.

[Meng et al., 2022] Meng, Y., Huang, J., Zhang, Y., and Han, J. (2022). Generating training data with language models: Towards zero-shot language understanding.

[Nguyen et al., 2020] Nguyen, D. T., Mummadi, C. K., Ngo, T. P. N., Nguyen, T. H. P., Beggel, L., and Brox, T. (2020). Self: Learning to filter noisy labels with self-ensembling. In *International Conference on Learning Representations*.

[Ouyang et al., 2022] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. (2022). Training language models to follow instructions with human feedback. In *Neural Information Processing Systems*.

[Ren et al., 2018] Ren, M., Zeng, W., Yang, B., and Urtasun, R. (2018). Learning to reweight examples for robust deep learning. In *International Conference on Machine Learning*.

[Shu et al., 2019] Shu, J., Xie, Q., Yi, L., Zhao, Q., Zhou, S., Xu, Z., and Meng, D. (2019). Meta-Weight-Net: Learning an explicit mapping for sample weighting. In *Neural Information Processing Systems*.

[Wei et al., 2022] Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., et al. (2022). Emergent abilities of large language models. *Transactions on Machine Learning Research*.