Selective Weak-to-Strong Generalization

Hao Lang, Fei Huang, Yongbin Li *

Tongyi Lab
{hao.lang, f.huang, shuide.lyb}@alibaba-inc.com

Abstract

Future superhuman models will surpass the ability of humans and humans will only be able to *weakly* supervise superhuman models. To alleviate the issue of lacking high-quality data for model alignment, some works on weak-to-strong generalization (W2SG) finetune a strong pretrained model with a weak supervisor so that it can generalize beyond weak supervision. However, the invariable use of weak supervision in existing methods exposes issues in robustness, with a proportion of weak labels proving harmful to models. In this paper, we propose a selective W2SG framework to avoid using weak supervision when unnecessary. We train a binary classifier P(IK) to identify questions that a strong model can answer and use its self-generated labels for alignment. We further refine weak labels with a graph smoothing method. Extensive experiments on three benchmarks show that our method consistently outperforms competitive baselines. Further analyses show that P(IK) can generalize across tasks and difficulties, which indicates selective W2SG can help superalignment.mple problems from the above-m

Introduction

Most existing AI alignment methods rely on the availability of human labelled data, such as human demonstrations for SFT (Wei et al. 2021), or human preferences for RLHF (Christiano et al. 2017; Ouyang et al. 2022). These methods have been used to build the most capable AI systems currently deployed (OpenAI 2023; Anthropic 2023).

However, future superhuman models will surpass the ability of humans in certain areas and behave in complex ways that humans cannot reliably evaluate (CAIS 2023). For example, if a superhuman model generates a billion lines of complicated code, humans will not be able to provide reliable supervision for superalignment (aligning superhuman models). In that case, the expected deficiency of human evaluation will limit the effectiveness of most alignment methods (Casper et al. 2023).

In response to the challenge of superalignment, recent work on weak-to-strong generalization (W2SG) has started to explore the potential of finetuning a strong pretrained model with a weak supervisor (Burns et al. 2023). In this study, a strong pretrained model is assumed to already have good representations of the alignment-relevant tasks, and a weak supervisor is expected to elicit what the strong model already knows. Empirical results show that the strong model can generalize beyond the weak supervision and even outperforms its weak supervisor on specific tasks given flawed training labels (Figure 1 top).

Despite their encouraging performance, existing W2SG approaches largely ignore to address a critical question:

- *Should we always use weak supervision to train the strong model?*

Our findings suggest that the answer is predominantly negative. First, in various tasks, the weak supervisor can only give incomplete or flawed training labels with noise. The desired generalization should be able to disagree with the weak supervision when the weak supervision is wrong (Burns et al. 2023). Second, existing W2SG approaches are still far from recovering the full capabilities of strong models. It suggests that alignment methods may scale poorly to superhuman models without additional work.

In this paper, we challenge the assumption of always using weak supervision by proposing a novel superalignment framework underpinned by a *selective weak-to-strong generalization* mechanism: the system proactively abstains from training a strong model with potentially detrimental weak labels (Figure 1 bottom). At the core of our framework is a strong model trained to estimate P(IK), the probability that it knows the answer to a question, on a given distribution (see Table 1 for an illustration) (Kadavath et al. 2022). The framework confidently uses labels generated by the strong model for alignment, when the strong model already knows the correct answers with high P(IK) scores.

Specifically, we train a strong model with a binary classification head to predict P(IK), which elicits the capability of evaluating its own knowledge state. We leverage additional existing datasets, generate strong model predictions, and obtain labels for classification by contrasting predictions with ground truth labels. Then, we jointly train a strong model to predict P(IK) and align with the optional weak supervision.

Furthermore, we refine weak supervision with labels generated by the strong model via a graph smoothing method (Luo et al. 2018; Li, Xiong, and Hoi 2021). Concretely, we divide questions into high P(IK) score IK questions with self-generated labels and low P(IK) score IDK

*Corresponding author.

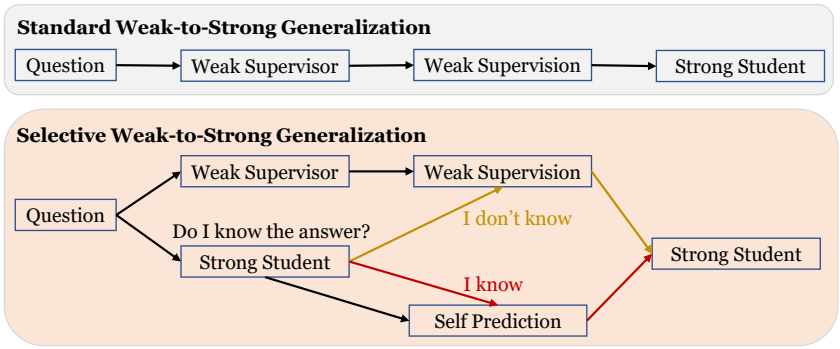

Figure 1: **Comparison of standard weak-to-strong generalization (W2SG) (top) and selective W2SG (bottom)**. Different from (top), which always uses weak supervision to train the strong model, we instead use the prediction of the strong model to train itself when the strong model knows the correct answer.

questions with weak labels, respectively. We construct a graph comprised of IK and IDK questions, and produce a graph-smoothed label for each IDK question by aggregating nearby nodes on the graph, based on the smoothness assumption (Van Engelen and Hoos 2020).

We perform comprehensive evaluations on three OpenAI weak-to-strong NLP benchmarks, which convert existing tasks to binary classification problems to make empirical progress in superalignment (Burns et al. 2023). Results show that our method achieves strong performance, outperforming methods that always use weak supervision.

Finally, we provide comprehensive analyses on the generalization ability of the P(IK) classifier. We show that P(IK) can generalize across tasks and difficulties, e.g., from science QA to reading comprehension, commonsense reasoning, and math. These analyses indicate that selective W2SG can help align future superhuman models.

Our major contributions are summarized:

1. We challenge the assumption of always using weak supervision for W2SG and propose a selective W2SG framework for superalignment.

2. We train a binary classifier P(IK) to identify whether using weak labels is necessary. We also refine weak labels with a graph smoothing method.

3. Extensive evaluations on three OpenAI weak-to-strong NLP benchmarks show that our approach outperforms competitive baselines. Further analyses indicate that the generalization ability of the P(IK) classifier is promising.

Related Work

AI alignment. AI alignment aims to steer already-capable models to behave in line with human values and intentions (Leike et al. 2018; Ji et al. 2023). Existing alignment methods finetune pretrained models using imitation learning on human demonstrations (Bain and Sammut 1995; Atkeson and Schaal 1997; Wei et al. 2021; Chung et al. 2024), reinforcement learning from human feedback (RLHF) (Christiano et al. 2017; Stiennon et al. 2020; Ouyang et al. 2022; Bai et al. 2022), or direct alignment algorithms like direct preference optimization (DPO) (Rafailov et al. 2024b,a). Both imitation learning and preference learning rely on high-quality supervision, a demand that becomes increasingly challenging as models become more capable than humans (Amodei et al. 2016).

Weak-to-strong generalization. Weak-to-strong generalization techniques seek to finetune a strong pretrained model to generalize well from weak supervision (Burns et al. 2023). This is typically pursued by assuming strong pretrained models should already have good representations of the alignment-relevant tasks, and thus we simply need a weak supervisor to elicit what the strong model already knows. Recently, Charikar, Pabbaraju, and Shiragur (2024); Mulgund and Pabbaraju (2025); Xu et al. (2025) theoretically quantify the gain in W2SG with misfit error.

A core question in this area is finding an effective policy to improve weak supervision. Burns et al. (2023) propose an auxiliary confidence loss to reinforce the strong model's confidence in its own predictions. Guo et al. (2024) propose an adaptively adjustable loss function for vision superalignment. Yao et al. (2025b,a) propose reverse KL and f-divergence losses to improve noise tolerance of the strong model. Lyu et al. (2024); Ye, Laidlaw, and Steinhardt (2025); Lang, Huang, and Li (2025) propose iterative updating methods, and Liu and Alahi (2024); Cui et al. (2024); Agrawal et al. (2024) propose ensemble methods. In comparison, this paper highlights the importance of understanding whether a strong model knows the correct answer (Kadavath et al. 2022) when improving weak supervision.

Semi-supervised learning. Our work is also related to semi-supervised learning (SSL) since they share the same *smoothness assumption*, i.e., samples close to each other are likely to receive similar labels (Luo et al. 2018; Van Engelen and Hoos 2020). The auxiliary confidence loss is also related to a key technique in SSL (Burns et al. 2023).

Preliminaries

We start by reviewing the weak-to-strong generalization pipeline in (Burns et al. 2023), which has also been adopted in subsequent work (Charikar, Pabbaraju, and Shiragur 2024; Guo et al. 2024). It usually consists of three steps:

Problem	P(IK)
Q: What disease occurs when the cell cycle is no longer regulated? A: cancer R:	99.6%
Q: Digestive enzymes are released, or secreted, by the organs of which body system? A: digestive system	99.6%
Q: Specific antigens on the surface of red blood cells determine what, which is important in cases of transfusion? A: blood type R:	99.6%
Q: What forms when nitrogen and sulfur oxides in air dissolve in rain? A: acid snow R:	13.4%
Q: Movements in the mantle cause the plates to move over time in a process called what? A: continental shift R:	13.2%
Q: What are the little sacs at the end of the bronchioles called? A: respiratory sacs R:	12.7%

Table 1: **Examples of P(IK) scores from a 14B model**. Harder problems have lower P(IK) scores. The binary classification problem is converted from the SciQ dataset. We also append the custom prompt *[There is a science knowledge question, followed by an answer. Respond with 1 if the answer is correct, and with 0 otherwise.]* to the problem. P(IK) refers to the probability that I Know the answer.

Performance	Method	SciQ		BoolQ		CosmosQA	
		Acc.	PGR	Acc.	PGR	Acc.	PGR
Weak performance		64.70		64.91		67.17	
Weak-to-strong performance	Finetune	77.90	45.99	65.89	3.92	71.99	21.37
	Finetune w/ auxi. loss	82.20	60.98	62.22	-10.76	73.37	27.49
	Finetune w/ prod. loss	81.60	58.89	61.02	-15.56	73.40	27.63
	Finetune w/ adap. loss	74.00	32.40	62.83	-8.32	67.77	2.66
	Finetune w/ rkl loss	78.50	48.08	66.29	5.52	74.10	30.73
	Finetune w/ js loss	80.00	53.31	65.80	3.56	71.96	21.24
	Ours	**83.30**	**64.81**	**68.64**	**14.92**	**75.11**	**35.21**
Strong ceiling performance		93.40		89.91		89.72	

Table 2: **Selective weak-to-strong generalization improves generalization**. Test accuarcy (%) and performance gap recovered (PGR) (%) of our approach and baselines on the binary classification tasks converted from NLP classification datasets. Here, we use **GPT-2** (Radford et al. 2019) for the weak model and **Qwen/Qwen-14B** (Bai et al. 2023) for the strong model, respectively. Accuracy of weak and strong models trained with ground truth are reported as weak performance and strong ceiling performance, respectively.

1. Create the weak supervisor. The first step is to create the weak supervisor by finetuning a small pretrained model on ground truth labels. The performance of the weak supervisor is called the *weak performance*.

2. Train a strong student model. The second step is to train a strong student model by finetuning a large pretrained model on weak labels generated by the weak supervisor. Its performance is called the *weak-to-strong performance*.

3. Train a strong ceiling model. The third step is to train a strong ceiling model by finetuning a large pretrained model on ground truth labels. This model's resulting performance is called the *strong ceiling performance*.

To measure the fraction of the performance gap that the strong student model can recover with weak supervision, Burns et al. (2023) define the performance gap recovered (PGR) using the above three performances:

$$\text{PGR} = \frac{\text{weak-to-strong} - \text{weak}}{\text{strong ceiling} - \text{weak}}.$$

To bridge gap between weak-to-strong performance and strong ceiling performance, they propose an auxiliary confidence loss that reinforces the strong model's confidence in its own predictions:

$$\mathcal{L}_{\text{conf}} = \text{CE}(f(x), (1-\alpha) \cdot f_w(x) + \alpha \cdot \hat{f}_t(x)),$$

where $\text{CE}(\cdot, \cdot)$ is the cross-entropy loss between two distributions on a given input x, $f(x) \in [0, 1]$ is the strong model predictive distribution, $f_w(x) \in [0, 1]$ is the weak label predictive distribution, $\hat{f}_t(x) = I[f(x) > t] \in \{0, 1\}$ is the hardened strong model prediction using a threshold t where I is the indicator function. α is a (fixed) weight (usually 0.5) to produce the cross-entropy target.

Methods

Overview

In this study, we build the strong student model following three steps: **1.** Train models to predict whether they can answer questions correctly; **2.** Estimate graph-smoothed weak labels for questions that a model cannot answer; **3.** Train a strong student model using graph-smoothed weak labels and self-generated labels selectively.

Performance	Method	SciQ		BoolQ		CosmosQA	
		Acc.	PGR	Acc.	PGR	Acc.	PGR
Weak performance		83.80		78.92		79.13	
Weak-to-strong performance	Finetune	88.10	44.79	80.85	17.56	81.37	21.15
	Finetune w/ auxi. loss	89.50	59.38	82.53	32.85	81.34	20.87
	Finetune w/ prod. loss	89.40	58.33	82.47	32.30	82.95	36.07
	Finetune w/ adap. loss	89.60	60.42	81.79	26.11	80.13	9.44
	Finetune w/ rkl loss	88.60	50.00	82.62	33.67	82.45	31.35
	Finetune w/ js loss	88.80	52.08	83.33	40.13	81.84	25.59
	Ours	**90.60**	**70.83**	**83.39**	**40.67**	**83.85**	**44.57**
Strong ceiling performance		93.40		89.91		89.72	

Table 3: **Selective weak-to-strong generalization improves generalization**. Test accuarcy (%) and performance gap recovered (PGR) (%) of our approach and baselines on the binary classification tasks converted from NLP classification datasets. Here, we use **Qwen/Qwen-1.8B** for the weak model and **Qwen/Qwen-14B** for the strong model, respectively (Bai et al. 2023). Accuracy of weak and strong models trained with ground truth are reported as weak performance and strong ceiling performance, respectively.

Training to Predict P(IK)

Central to our framework is the idea of *selective weak-to-strong generalization*, where the system decides whether the strong student model could benefit from weak supervision and abstains from training with weak labels when it is deemed unnecessary (Figure 1 bottom).

Concretely, after observing the current question, the strong model explicitly predicts whether or not it can correctly answer the question. If the strong model knows the correct answer, we train the strong model with labels generated by itself; otherwise, we still train the strong model with weak labels.

This approach is inspired by the findings in Kadavath et al. (2022). They show that strong pretrained models can be trained to predict whether they know the answer to any given question, denoting the probability they assign as P(IK) (for Probability that I Know the answer). This is fundamentally a question about the strong models themselves, which demonstrate their abilities in directly evaluating their own knowledge state.

To effectively train P(IK), we create a training set in the form of (few-shot prompt + question, ground truth label). There are just two choices for the label (IK / IDK), where IK means the model knows the correct answer, and IDK means the model does not know. For a given question Q, if the model's sampled answer is correct, our training set contains a datapoint (Q, IK); otherwise, a datapoint (Q, IDK). We use a 4-shot prompt simply to ensure that the model's sampled answers are in the correct format.

We implement a binary classifier P(IK) by equipping the model with a linear classification head with two outputs. During P(IK) training, we finetune the entire model along with the head using a cross-entropy loss $\mathcal{L}_{\text{ik}}$. Table 1 shows some examples of P(IK) scores from a 14 billion parameter model on a few example questions where the model sensibly should or should not know the answers.

As a note, in a later section we study out-of-distribution generalization and easy-to-hard generalization of P(IK). We believe that generalization of P(IK) is crucial for selective weak-to-strong generalization because it will be applied in future novel and complex tasks.

Estimating Graph-smoothed Weak Labels

To further capitalize on the ability of a strong model in predicting P(IK), we refine weak labels with a graph smoothing method (Luo et al. 2018; Lang et al. 2022).

Given predictions of P(IK) in the current batch, questions to train the strong model can be divided into two groups: IK questions $\mathcal{D}_{\text{ik}}$ with P(IK) scores above a threshold γ and IDK questions $\mathcal{D}_{\text{idk}}$ with P(IK) scores below γ. Note that IK questions have reliable labels generated by the strong model, and IDK questions have unreliable weak labels provided by the weak supervisor.

We construct a fully connected unidirectional embedding graph $\mathcal{G}$ using samples in $\mathcal{D} = \mathcal{D}_{\text{ik}} \cup \mathcal{D}_{\text{idk}}$. We first map each sample $x \in \mathcal{D}$ into an embedding z (computed from the final transformer layer), and then use all these embeddings as nodes for $\mathcal{G}$. We also assign a prior label $l_p(x)$ to each sample $x \in \mathcal{D}$ to represent its annotation, i.e., for a sample $x \in \mathcal{D}_{\text{ik}}$, $l_p(x)$ is defined as the self generated label by the strong model, and for a sample $x \in \mathcal{D}_{\text{idk}}$, $l_p(x)$ is defined as the corresponding weak label.

For each sample $x \in \mathcal{D}_{\text{idk}}$, a graph-smoothed label $l_g(x)$ is obtained by aggregating adjacent nodes on $\mathcal{G}$. Specifically, to conform to the smoothness assumption, we try to minimize the following distance when determining $l_g(x)$:

$$\begin{aligned} &\alpha \cdot d[l_g(x), l_p(x)] + (1-\alpha) \sum_{x_j \in \mathcal{D}} a_j \cdot d[l_g(x), l_p(x_j)] \\ &a_j = \frac{\exp(z \cdot z_j/\tau)}{\sum_{k=1}^{|\mathcal{D}|} \exp(z \cdot z_k/\tau)}, \end{aligned} \tag{1}$$

where $0 \leq \alpha \leq 1$ is a weight, d is a distance function, $\tau > 0$ is a temperature. The second term in Eq. 1 enforces the smoothness assumption by encouraging $l_g(x)$ to have similar labels with its nearby samples, whereas the first term tries to maintain $l_g(x)$ to meet its original annotation $l_p(x)$. For simplicity, we implement d as the Euclidean distance

Method	SciQ		BoolQ		CosmosQA	
	Acc.	PGR	Acc.	PGR	Acc.	PGR
Ours w/o IK	89.86	63.13	82.28	30.57	81.36	21.06
Ours w/o GS	90.23	66.98	83.11	38.13	83.12	37.68
MTL	89.98	64.38	82.56	33.12	82.27	29.65
Ours	**90.60**	**70.83**	**83.39**	**40.67**	**83.85**	**44.57**

Table 4: Ablation on main components of our model.

Method	SciQ	BoolQ
Finetune	88.10	80.85
Ours (+CosmosQA)	**90.70**	83.20
Ours (+Winograde)	89.80	82.99
Ours (+ARC)	90.60	**83.39**

Table 5: **Ablation on the dataset to train the P(IK) classifier**. The performance is evaluated with accuracy.

here, and thus minimizing Eq. 1 yields:

$$l_g(x) = \alpha \cdot l_p(x) + (1-\alpha) \sum_{x_j \in \mathcal{D}} a_j \cdot l_p(x_j) \tag{2}$$

Note that the result we derived in Eq. 2 follows most previous graph-smoothing methods in semi-supervised learning (Van Engelen and Hoos 2020).

Training Models using Labels Selectively

After producing the graph-smoothed weak labels, we construct a dataset with improved labels:

$$\mathcal{M} = \{(x_i, l_p(x_i)) \cup (x_j, l_g(x_j)) | x_i \in \mathcal{D}_{\text{ik}}, x_j \in \mathcal{D}_{\text{idk}}\} \tag{3}$$

We can finetune a model on $\mathcal{M}$ with a cross-entropy loss $\mathcal{L}_{\text{gen}}$. The final training loss for finetuning a large pretrained student model is:

$$\mathcal{L} = \mathcal{L}_{\text{gen}} + \lambda \cdot \mathcal{L}_{\text{ik}}, \tag{4}$$

where $\mathcal{L}_{\text{gen}}$ is optimized on a dataset to evaluate weak-to-strong performance, and $\mathcal{L}_{\text{ik}}$ is optimized on another dataset to train P(IK), λ is a weight.

Experiments

Tasks

We adopt the evaluation protocol of prior work (Burns et al. 2023), and conduct experiments in NLP tasks on three benchmark datasets: SciQ (Welbl, Liu, and Gardner 2017), BoolQ (Clark et al. 2019), and CosmosQA (Huang et al. 2019). We convert each dataset to a binary classification problem. For multiple-choice datasets, given a data point with a question Q and k candidate answers A, we construct k new data points of the form (Q, A_i), where the label is 1 for the correct answers and 0 for all the incorrect answers. We also keep the same number of correct and incorrect answers per question to maintain class balance.

Experimental Setups and Metrics

Following (Burns et al. 2023), we randomly sample at most 20k data points from each task and split them in half. We train a weak model on the first half of the data points and use its prediction on the other half as the weak labels. The weak labels are soft labels (Hinton, Vinyals, and Dean 2015). We report the accuracy and performance gap recovered (PGR) of the strong student model on the test set in all tasks.

Implementation Details

Our implementations of data preprocessing, weak and strong model training are based on the OpenAI weak-to-strong codebase and its default hyper-parameters (Burns et al. 2023). Specifically, we use Qwen/Qwen-14B (Bai et al. 2023) as the large pretrained model for training strong models. Meanwhile, we use GPT-2 (Radford et al. 2019) and Qwen/Qwen-1.8B as two small pretrained models for training weak models, which have different gaps in compute between weak and strong models.

In order to adapt weak and strong models to the converted binary classification setting, we equip each model with another linear classification head with two outputs. We use $\lambda = 1$, $\gamma = 0.8$, $\alpha = 0.9$, and $\tau = 0.1$ in all experiments. We train all models for two epochs with a batch size of 32. We conduct all experiments on a single 8×A100 machine.

Baselines

We compare our approach with competitive baseline approaches: 1. **Finetune** (Burns et al. 2023) naively finetunes strong pretrained models on labels generated by a weak model; 2. **Finetune w/ auxi. loss** (Burns et al. 2023) finetunes strong models with an auxiliary confidence loss, which reinforces the strong model's confidence in its own predictions; 3. **Finetune w/ prod. loss** (Burns et al. 2023) finetunes strong models with a confidence-like loss which sets the cross entropy targets to the product of weak labels and strong model predictions. 4. **Finetune w/ adap. loss** (Guo et al. 2024) finetunes strong models with an adaptively adjustable loss using the discrepancy between the soft label and the hard label. 5. **Finetune w/ rkl loss** (Yao et al. 2025b) finetunes strong models with reverse KL divergence loss. 6. **Finetune w/ js loss** (Yao et al. 2025a) finetunes strong models with Jensen–Shannon divergence loss. We also report the **weak performance** and the **strong ceiling performance** defined in the preliminaries section. Note that the strong ceiling performance is regarded as the upper bound of the **weak-to-strong performance** when only weak labels are considered.

	In-Dist Generalization		OOD Generalization	
	Training Data	AUROC	Training Data	AUROC
SciQ	SciQ	92.14	ARC	89.34
BoolQ	BoolQ	85.08	ARC	79.02
CosmosQA	CosmosQA	88.01	ARC	79.92
Winograde	Winograde	73.43	ARC	63.55
GSM8K	GSM8K	69.27	ARC	47.73

Table 6: **Overview comparing out-of-distribution generalization to in-distribution generalization performance of P(IK)**. All AUROC scores (%) are computed using the Qwen/Qwen-14B P(IK) classifiers. Even when we only train on ARC, we see decent generalization to other tasks.

Main Results

Table 2 and Table 3 show the results of each approach on the binary classification tasks converted from SciQ, BoolQ, and CosmosQA datasets. Here, our approach trains the binary classier P(IK) on questions from the ARC dataset (Clark et al. 2018). We use GPT-2 for the weak model in Table 2 and use Qwen/Qwen-1.8B for the weak model in Table 3. In each task, we observe that PGRs of strong student models naively finetuned on weak labels are all positive, which indicates promising weak-to-strong generalization.

At the same time, we find that our approach significantly outperforms each strong student baseline, including the naive baseline finetuned on weak labels or more sophisticated baselines equipped with a confidence loss term on all three tasks. Compared with the promising baseline Finetune w/ auxi. loss, our approach brings up from a PGR of 60.98% to 64.81% in SciQ, -10.76% to 14.92% in BoolQ, and 27.49% to 35.21% in CosmosQA, when using weak model GPT-2, and brings up from a PGR of 59.38% to 70.83% in SciQ, 32.85% to 40.67% in BoolQ, and 20.87% to 44.57% in CosmosQA, when using weak model Qwen/Qwen-1.8B. Our approach also obtains the best test accuracy among all compared strong students. The performance gain shows the advantage of selective W2SG, which helps elicit the capabilities of the strong model with weak supervision.

Ablation Studies

We provide comprehensive ablation studies to understand the efficacy of selective weak-to-strong generalization framework. All ablations are conducted using Qwen/Qwen-1.8B for the weak model and Qwen/Qwen-14B for the strong model.

Ablation on model main components. We verify the effect of each component in our selective framework by testing following variants: 1. **Ours w/o IK** removes training the P(IK) classier. In this variant, IDK questions $\mathcal{D}_{\text{idk}}$ contains all the questions to train the student model, and the loss shown in Eq. 4 is optimized by setting $\lambda = 0$; 2. **Ours w/o GS** removes estimating the graph-smoothed weak labels. In this variant, $l_g(x)$ of each sample $x \in \mathcal{D}_{\text{idk}}$ shown in Eq. 3 is replaced with the prior label $l_p(x)$. 3. **MTL** implements multi-task learning with one head for P(IK) and another head for alignment. The head for P(IK) is trained with ground truth labels from the ARC dataset while the other head is trained with weak labels.

Results in Table 4 indicate that our method outperforms all ablation models in terms of test accuracy and PGR. We can further observe that: **1.** Training the P(IK) classier to identify IK questions brings the largest improvement compared to other components. This proves the importance of understanding whether a strong model knows the correct answer to a question for selective W2SG. **2.** Naively training the P(IK) classier without selecting labels for alignment degenerates the model performance by a large margin. This shows the effectiveness of improved label quality produced by our selective framework.

Ablation on the dataset to train the P(IK) classifier. We analyze the impact of the dataset to train the P(IK) classifier on the final performance. In Table 5, we train the P(IK) classifier on questions from three different datasets: CosmosQA, Winograde (Sakaguchi et al. 2021), and ARC. We can observe that our selective W2SG method consistently outperforms methods that always use weak supervision, when training the P(IK) classifier with different extra datasets. It also suggests that strong models can be elicited to evaluate their own knowledge state. We present analyses of the P(IK) generalization ability in the next subsection.

Further Analysis

Achieving good generalization of P(IK) is important for selective W2SG, since it will be applied in future novel and complex tasks for superalignment. We are interested in studying the generalization of P(IK) across various dimensions, i.e., cross-task, cross-domain, and cross-difficulty.

Specifically, we train P(IK) classifiers only on ARC and then evaluating on SciQ, BoolQ, CosmosQA, Winograde, and GSM8K (Cobbe et al. 2021). These datasets cover diverse domains such as science QA, reading comprehension, commonsense reasoning, and math.Meanwhile, these datasets also cover diverse difficulties, where Winograde and GSM8K are much harder than the other datasets, with relatively lower test accuracy. Hence, evaluation on these datasets help us explore out-of-distribution generalization and easy-to-hard generalization (Sun et al. 2024) of P(IK). In this analysis, we implement a variant of our model that only optimizes the loss $\mathcal{L}_{\text{ik}}$ in Eq. 4, i.e., $\mathcal{L}_{\text{gen}}$ is removed.

Table 6 gives an overview of generalization performance for the P(IK) classifier that is trained on ARC, compared to its in-distribution performance. Figure 2 gives a detailed view of how the distribution of P(IK) changes depending on

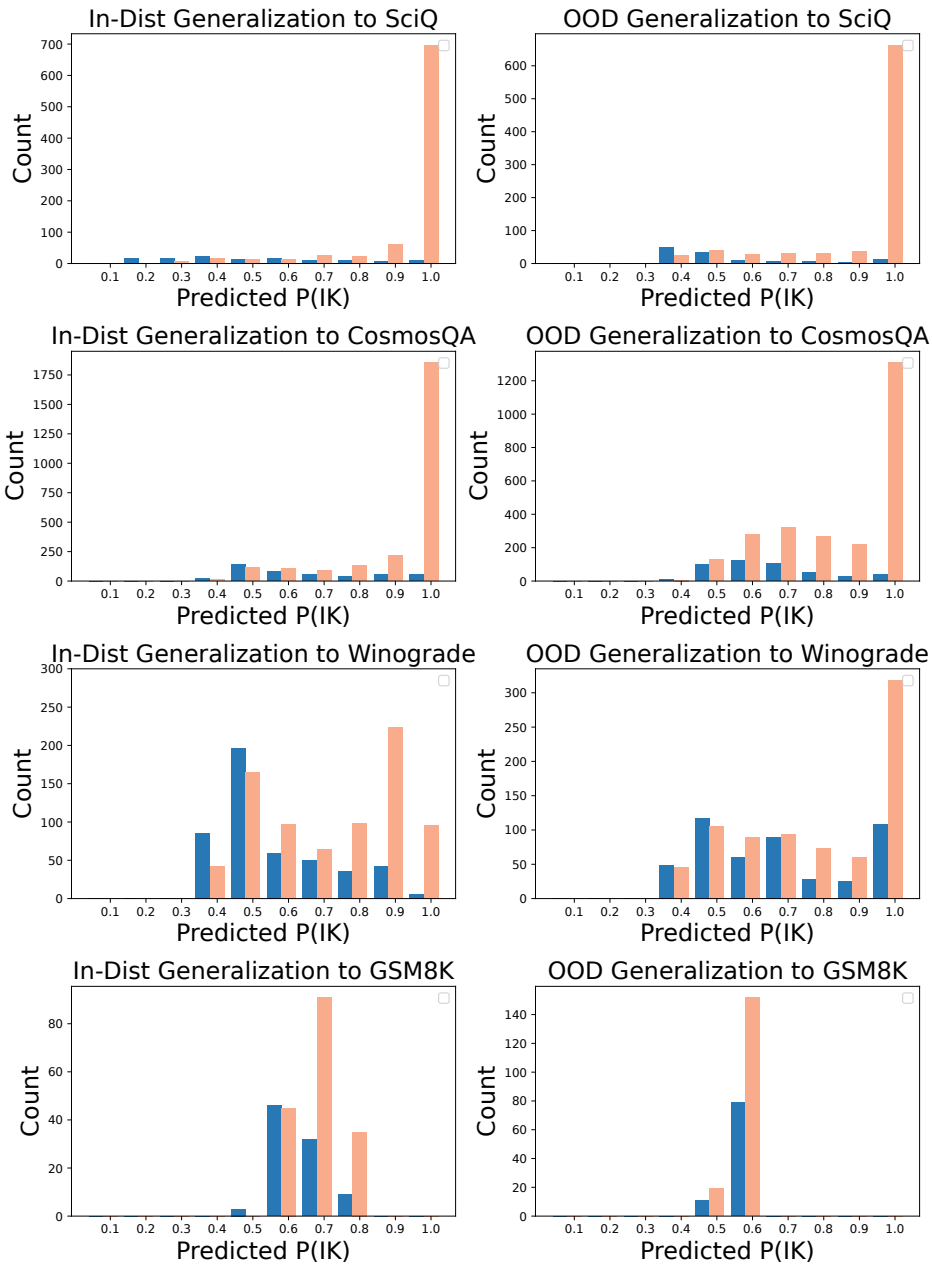

Figure 2: **Generalization of P(IK)**. The left side of this figure includes distributions of P(IK) from a Qwen/Qwen-14B P(IK) classifier that was trained on in-distribution data, i.e., SciQ, CosmosQA, Winograde, and GSM8K, respectively. The right side includes distributions of P(IK) that was trained on just ARC. Blue bar represents P(IK)=0, and orange bar represents P(IK)=1.

training data. We find that strong models do exhibit a degree of generalization of P(IK) across tasks and difficulties.

Limitations and Conclusion

Limitations. Although our proposed method is found to be effective in all our experiments, the difference between strong and weak models is only in the size of pretrained models in our setup. However, in the future, strong models may also differ in reasoning and planning abilities. Furthermore, since weak-to-strong deception phenomenon exists (Yang et al. 2024), the P(IK) classifier might not be perfectly robust for future superalignment. It highlights the urgent need to pay more attention to alignment reliability.

Conclusion. In this paper, we challenge the assumption of always using weak supervision for W2SG. In response, we propose a selective framework, where strong models identify whether using weak labels is necessary. Extensive evaluations show that our approach outperforms SOTA baselines.

References

Agrawal, A.; Ding, M.; Che, Z.; Deng, C.; Satheesh, A.; Langford, J.; and Huang, F. 2024. EnsemW2S: Can an Ensemble of LLMs be Leveraged to Obtain a Stronger LLM? *arXiv preprint arXiv:2410.04571*.

Amodei, D.; Olah, C.; Steinhardt, J.; Christiano, P.; Schulman, J.; and Mané, D. 2016. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*.

Anthropic. 2023. Introducing claude. https://www.anthropic.com/index/introducing-claude.

Atkeson, C. G.; and Schaal, S. 1997. Robot learning from demonstration. In *ICML*, volume 97, 12–20.

Bai, J.; Bai, S.; Chu, Y.; Cui, Z.; Dang, K.; Deng, X.; Fan, Y.; Ge, W.; Han, Y.; Huang, F.; et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Bai, Y.; Jones, A.; Ndousse, K.; Askell, A.; Chen, A.; DasSarma, N.; Drain, D.; Fort, S.; Ganguli, D.; Henighan, T.; et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Bain, M.; and Sammut, C. 1995. A Framework for Behavioural Cloning. In *Machine Intelligence 15*, 103–129.

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*.

CAIS. 2023. Statement on AI Risk. https://www.safe.ai/work/statement-on-ai-risk.

Casper, S.; Davies, X.; Shi, C.; Gilbert, T. K.; Scheurer, J.; Rando, J.; Freedman, R.; Korbak, T.; Lindner, D.; Freire, P.; et al. 2023. Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217*.

Charikar, M.; Pabbaraju, C.; and Shiragur, K. 2024. Quantifying the Gain in Weak-to-Strong Generalization. *arXiv preprint arXiv:2405.15116*.

Christiano, P. F.; Leike, J.; Brown, T.; Martic, M.; Legg, S.; and Amodei, D. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Chung, H. W.; Hou, L.; Longpre, S.; Zoph, B.; Tay, Y.; Fedus, W.; Li, Y.; Wang, X.; Dehghani, M.; Brahma, S.; et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70): 1–53.

Clark, C.; Lee, K.; Chang, M.-W.; Kwiatkowski, T.; Collins, M.; and Toutanova, K. 2019. BoolQ: Exploring the surprising difficulty of natural yes/no questions. *arXiv preprint arXiv:1905.10044*.

Clark, P.; Cowhey, I.; Etzioni, O.; Khot, T.; Sabharwal, A.; Schoenick, C.; and Tafjord, O. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Cobbe, K.; Kosaraju, V.; Bavarian, M.; Chen, M.; Jun, H.; Kaiser, L.; Plappert, M.; Tworek, J.; Hilton, J.; Nakano, R.; et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Cui, Z.; Zhang, Z.; Sun, G.; Wu, W.; and Zhang, C. 2024. Bayesian weaks-to-strong from text classification to generation. *arXiv preprint arXiv:2406.03199*.

Guo, J.; Chen, H.; Wang, C.; Han, K.; Xu, C.; and Wang, Y. 2024. Vision superalignment: Weak-to-strong generalization for vision foundation models. *arXiv preprint arXiv:2402.03749*.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Huang, L.; Bras, R. L.; Bhagavatula, C.; and Choi, Y. 2019. Cosmos QA: Machine reading comprehension with contextual commonsense reasoning. *arXiv preprint arXiv:1909.00277*.

Ji, J.; Qiu, T.; Chen, B.; Zhang, B.; Lou, H.; Wang, K.; Duan, Y.; He, Z.; Zhou, J.; Zhang, Z.; et al. 2023. Ai alignment: A comprehensive survey. *arXiv preprint arXiv:2310.19852*.

Kadavath, S.; Conerly, T.; Askell, A.; Henighan, T.; Drain, D.; Perez, E.; Schiefer, N.; Dodds, Z. H.; DasSarma, N.; Tran-Johnson, E.; et al. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Lang, H.; Huang, F.; and Li, Y. 2025. Debate Helps Weak-to-Strong Generalization. *arXiv preprint arXiv:2501.13124*.

Lang, H.; Zheng, Y.; Sun, J.; Huang, F.; Si, L.; and Li, Y. 2022. Estimating soft labels for out-of-domain intent detection. *arXiv preprint arXiv:2211.05561*.

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

Li, J.; Xiong, C.; and Hoi, S. C. 2021. Comatch: Semi-supervised learning with contrastive graph regularization. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9475–9484.

Liu, Y.; and Alahi, A. 2024. Co-supervised learning: Improving weak-to-strong generalization with hierarchical mixture of experts. *arXiv preprint arXiv:2402.15505*.

Luo, Y.; Zhu, J.; Li, M.; Ren, Y.; and Zhang, B. 2018. Smooth neighbors on teacher graphs for semi-supervised learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8896–8905.

Lyu, Y.; Yan, L.; Wang, Z.; Yin, D.; Ren, P.; de Rijke, M.; and Ren, Z. 2024. MACPO: weak-to-strong alignment via multi-agent contrastive preference optimization. *arXiv preprint arXiv:2410.07672*.

Mulgund, A.; and Pabbaraju, C. 2025. Relating misfit to gain in weak-to-strong generalization beyond the squared loss. *arXiv preprint arXiv:2501.19105*.

OpenAI. 2023. Gpt-4 technical report. https://openai.com/index/gpt-4-research/.

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. *Advances in neural information processing systems*, 35: 27730–27744.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8): 9.

Rafailov, R.; Chittepu, Y.; Park, R.; Sikchi, H.; Hejna, J.; Knox, B.; Finn, C.; and Niekum, S. 2024a. Scaling laws for reward model overoptimization in direct alignment algorithms. *arXiv preprint arXiv:2406.02900*.

Rafailov, R.; Sharma, A.; Mitchell, E.; Manning, C. D.; Ermon, S.; and Finn, C. 2024b. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Sakaguchi, K.; Bras, R. L.; Bhagavatula, C.; and Choi, Y. 2021. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9): 99–106.

Stiennon, N.; Ouyang, L.; Wu, J.; Ziegler, D.; Lowe, R.; Voss, C.; Radford, A.; Amodei, D.; and Christiano, P. F. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33: 3008–3021.

Sun, Z.; Yu, L.; Shen, Y.; Liu, W.; Yang, Y.; Welleck, S.; and Gan, C. 2024. Easy-to-hard generalization: Scalable alignment beyond human supervision. *arXiv preprint arXiv:2403.09472*.

Van Engelen, J. E.; and Hoos, H. H. 2020. A survey on semi-supervised learning. *Machine Learning*, 109(2): 373–440.

Wei, J.; Bosma, M.; Zhao, V. Y.; Guu, K.; Yu, A. W.; Lester, B.; Du, N.; Dai, A. M.; and Le, Q. V. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Welbl, J.; Liu, N. F.; and Gardner, M. 2017. Crowdsourcing multiple choice science questions. *arXiv preprint arXiv:1707.06209*.

Xu, G.; Yao, W.; Wang, Z.; and Liu, Y. 2025. On the emergence of weak-to-strong generalization: A bias-variance perspective. *arXiv preprint arXiv:2505.24313*.

Yang, W.; Shen, S.; Shen, G.; Yao, W.; Liu, Y.; Gong, Z.; Lin, Y.; and Wen, J.-R. 2024. Super (ficial)-alignment: Strong models may deceive weak models in weak-to-strong generalization. *arXiv preprint arXiv:2406.11431*.

Yao, W.; Xu, G.; Tang, H.; Yang, W.; Di, D.; Wang, Z.; and Liu, Y. 2025a. On Weak-to-Strong Generalization and f-Divergence. *arXiv preprint arXiv:2506.03109*.

Yao, W.; Yang, W.; Wang, Z.; Lin, Y.; and Liu, Y. 2025b. Revisiting weak-to-strong generalization in theory and practice: Reverse kl vs. forward kl. *arXiv preprint arXiv:2502.11107*.

Ye, Y.; Laidlaw, C.; and Steinhardt, J. 2025. Iterative label refinement matters more than preference optimization under weak supervision. *arXiv preprint arXiv:2501.07886*.