Google DeepMind

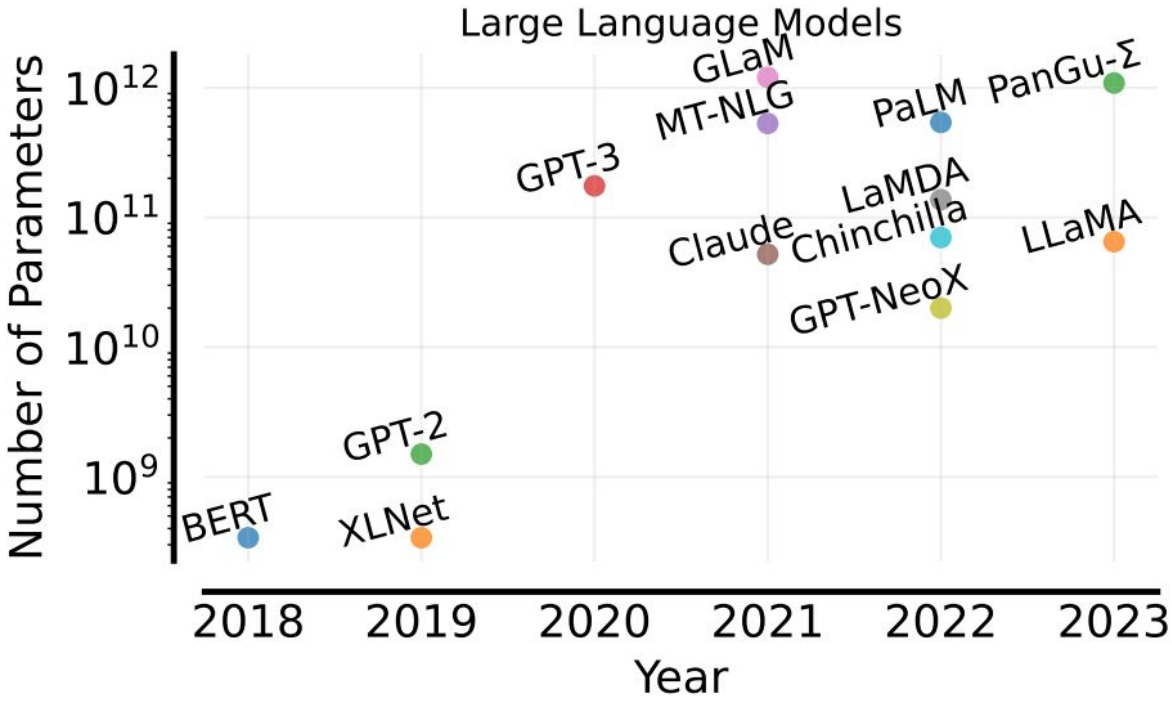
How to Distill Your Autoregressive Model

Rishabh Agarwal @ DLCT

August 2023

arxiv.org/abs/2306.13649

ML models are quite large ..



Aren’t bigger models always better?

- Deployment of “large” models limited by either their **inference cost** or **memory footprint**.
  - You can’t put PaLM 540B on your smartphone.
  - You don’t want to typically wait several minutes for an ML model to generate an output.

So what? Model compression.

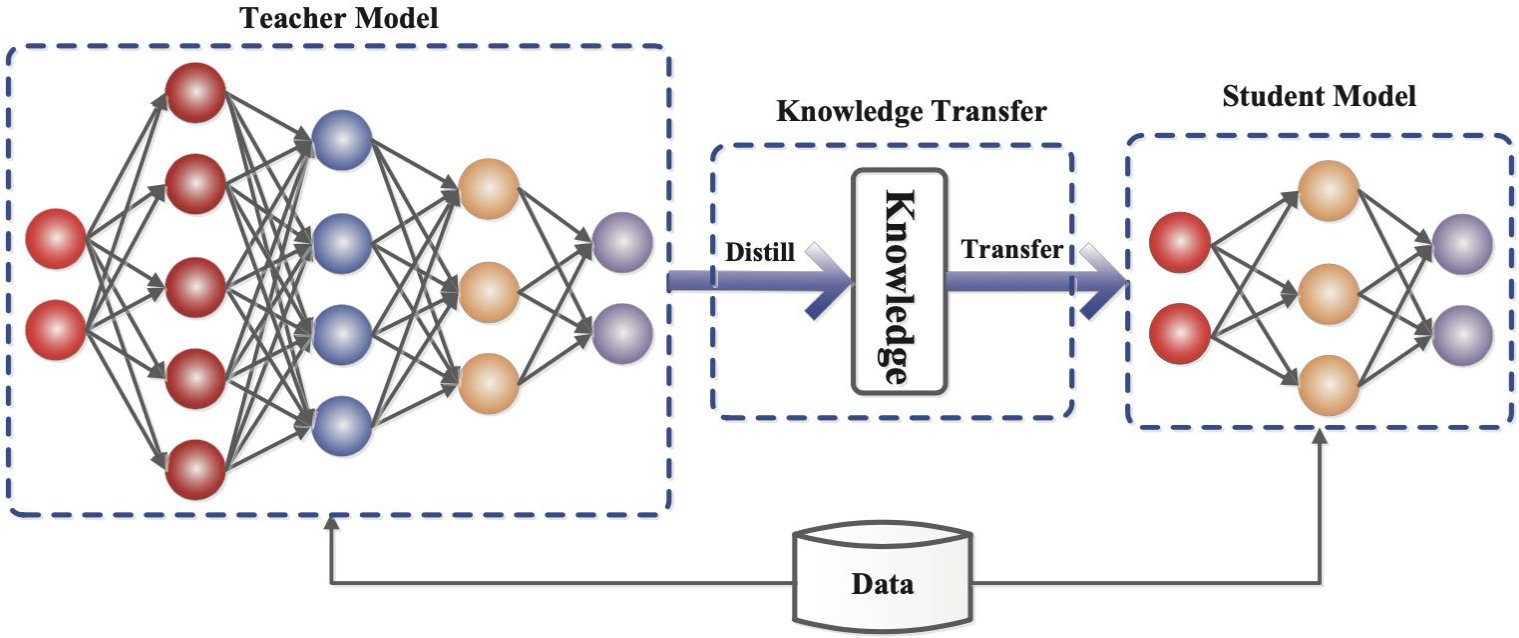
What is Model Compression?

The main idea is to simplify the model without diminishing accuracy. A simplified model means reduced in size and/or latency from the original.

- ➢ Size reduction can be achieved by reducing the model parameters and thus using less RAM.

- ➢ Latency reduction can be achieved by decreasing the time it takes for the model to make a prediction, and thus lowering energy consumption at runtime (and carbon footprint).

Knowledge Distillation (KD)



The generic framework of teacher-student knowledge distillation training. (Image source: Gou et al. 2020)

Current KD methods are suboptimal for autoregressive models!

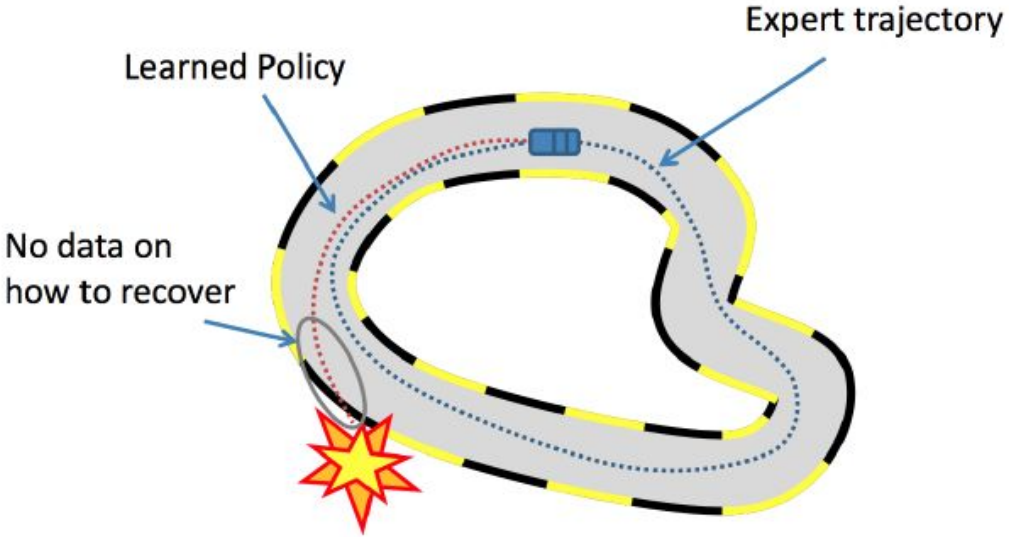
$$L_{SD}(\theta) := \mathbb{E}_{(x,y)\sim(X,Y)}\left[\mathcal{D}_{KL}\left(p_{\mathrm{T}} \| p_{\mathrm{S}}^{\theta}\right)(y|x)\right]$$

Informal: KD methods were originally designed for **single-step** classification / regression but auto-regressive models generate **multi-step** output sequentially token by token.

For people with RL background: Autoregressive models can be thought of as “agents” and current KD methods are analogous to behavior cloning.

Distribution Mismatch (Exposure Bias)

Existing KD methods typically train on a fixed dataset of output sequences. This results in a mismatch with the sequences generated by the student auto-regressively during deployment.

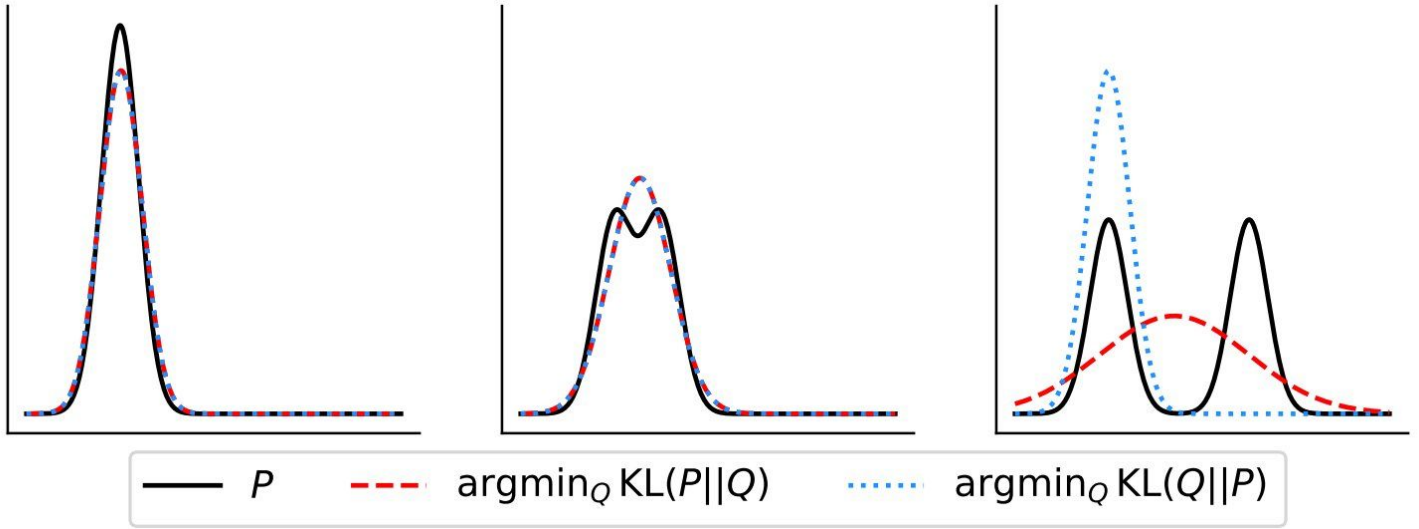


Well-known in the Imitation learning community.

Model Underspecification

Common KD objective is to maximize the likelihood of samples likely under the teacher.

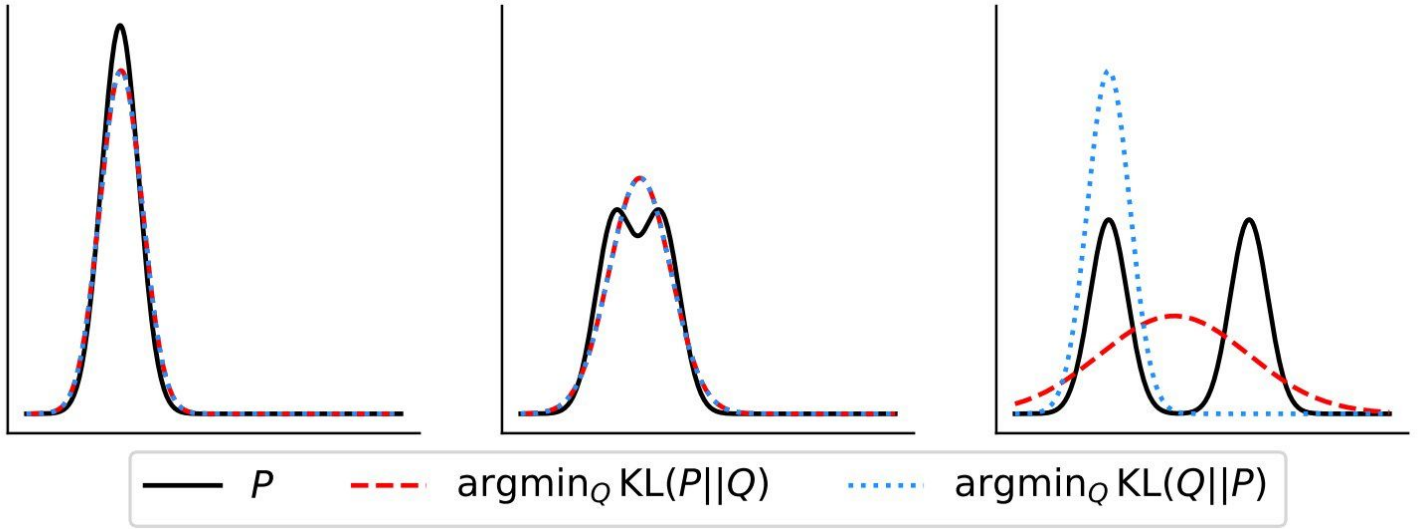
However, the student is often not expressive enough to fit the teacher's distribution and MLE can lead to unnatural student-generated samples. MLE = KL(P||Q).



Model Underspecification

Common KD objective is to maximize the likelihood of samples likely under the teacher.

However, the student is often not expressive enough to fit the teacher's distribution and MLE can lead to unnatural student-generated samples. MLE = KL(P||Q).



Our Solution: GKD

- GKD mitigates distribution mismatch by sampling output sequences from the student during training.
- GKD handles model under-specification by optimizing alternative divergences, such as reverse KL, that focus on generating samples from the student that are likely under the teacher's distribution

$$L_{\mathrm{GKD}}(\theta) := (1-\lambda)\,\mathbb{E}_{(x,y)\sim(X,Y)}\left[\mathcal{D}(p_{\mathrm{T}}\|p_{\mathrm{S}}^{\theta})(y|x)\right] + \lambda\,\mathbb{E}_{x\sim X}\left[\mathbb{E}_{y\sim p_{\mathrm{S}}(\cdot|x)}\left[\mathcal{D}(p_{\mathrm{T}}\|p_{\mathrm{S}}^{\theta})(y|x)\right]\right]$$

GKD: Algorithm in a box

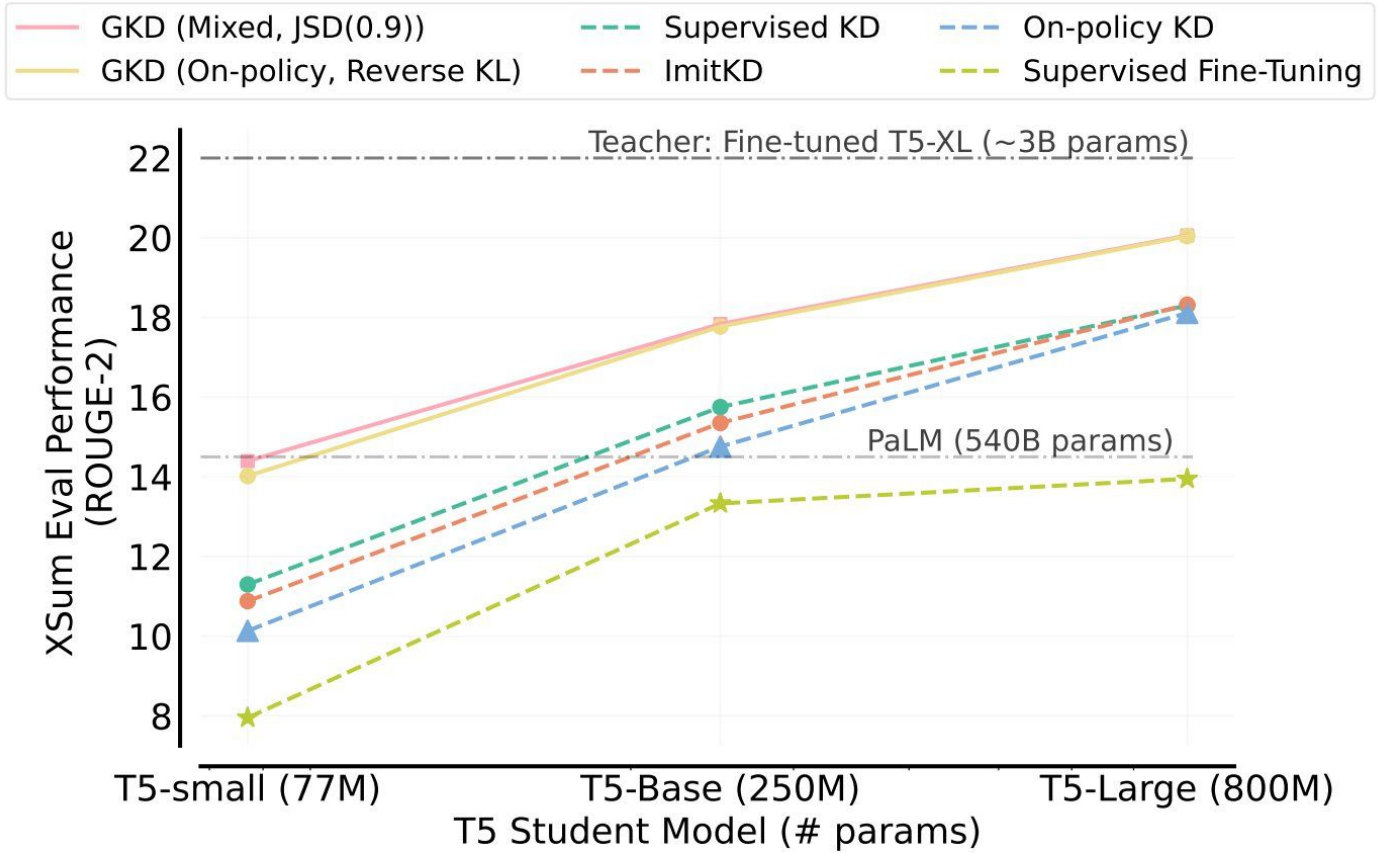
Algorithm 1 Generalized Knowledge Distillation (GKD)

Given: Teacher model p_{T}, Student Model p_{S}^{θ}, Dataset (X, Y) contain input contexts x and possibly output sequences y.
Hyperparameters: Student data fraction $\lambda \in [0, 1]$, Divergence $\mathcal{D}$, Learning rate η
for each step $k = 1, \ldots, K$ **do**
  Generate a random value $u \sim Uniform(0, 1)$
  if $u \leq \lambda$ **then**
    Sample batch of contexts from X and for each x, sample one output $y \sim p_{\mathrm{S}}^{\theta}(\cdot|x)$ to obtain $B = \{(x_b, y_b)\}_{b=1}^{B}$
  else
    Sample batch of contexts and outputs from (X, Y) to obtain $B = \{(x_b, y_b)\}_{b=1}^{B}$.
  Update θ to minimize L_{GKD}: $\theta \leftarrow \theta - \eta \frac{1}{B} \sum_{(x,y) \in B} \nabla_{\theta} \mathcal{D}(p_{\mathrm{T}} \| p_{\mathrm{S}}^{\theta})(y|x)$

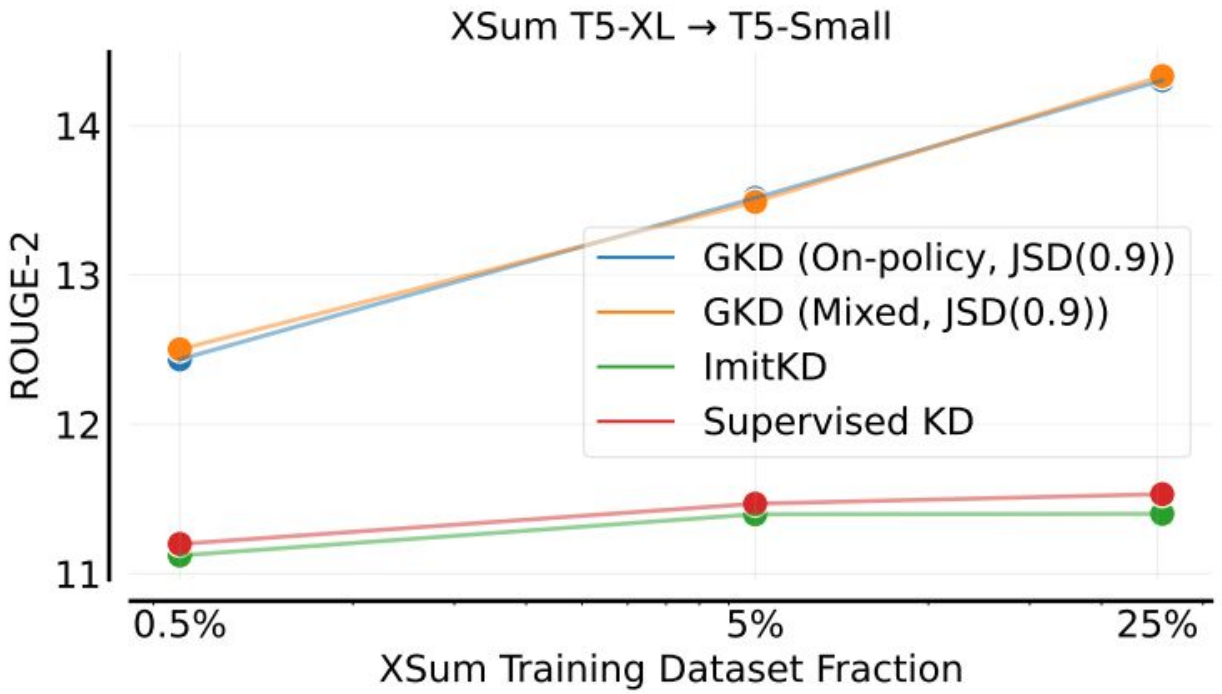
Experiments: Baselines

- **Supervised FT**: See Section 4 It directly fine-tunes the student on either ground-truth or teacher-generated outputs.
- **Supervised KD**: See Eq (4), GKD with $\lambda = 0$ and forward KL as divergence $\mathcal{D}$.
- **On-policy KD**: See Eq (5), GKD with $\lambda = 1$ and forward KL as divergence. Note that purely on-policy distillation hasn't been employed for auto-regressive models.
- **ImitKD** (Lin et al., 2020): GKD with $\lambda = 0.5$ and forward KL as divergence.

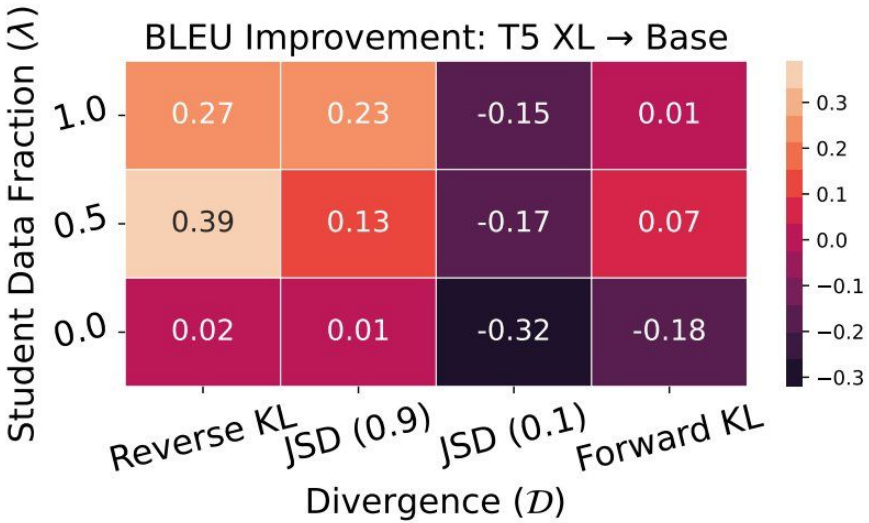
Results: News Summarization (XSum)

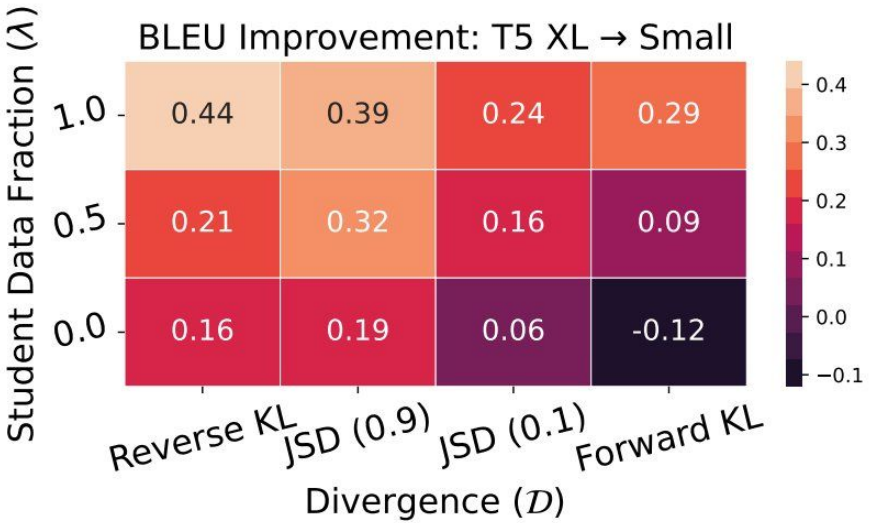


Data Scaling: News Summarization (XSum)

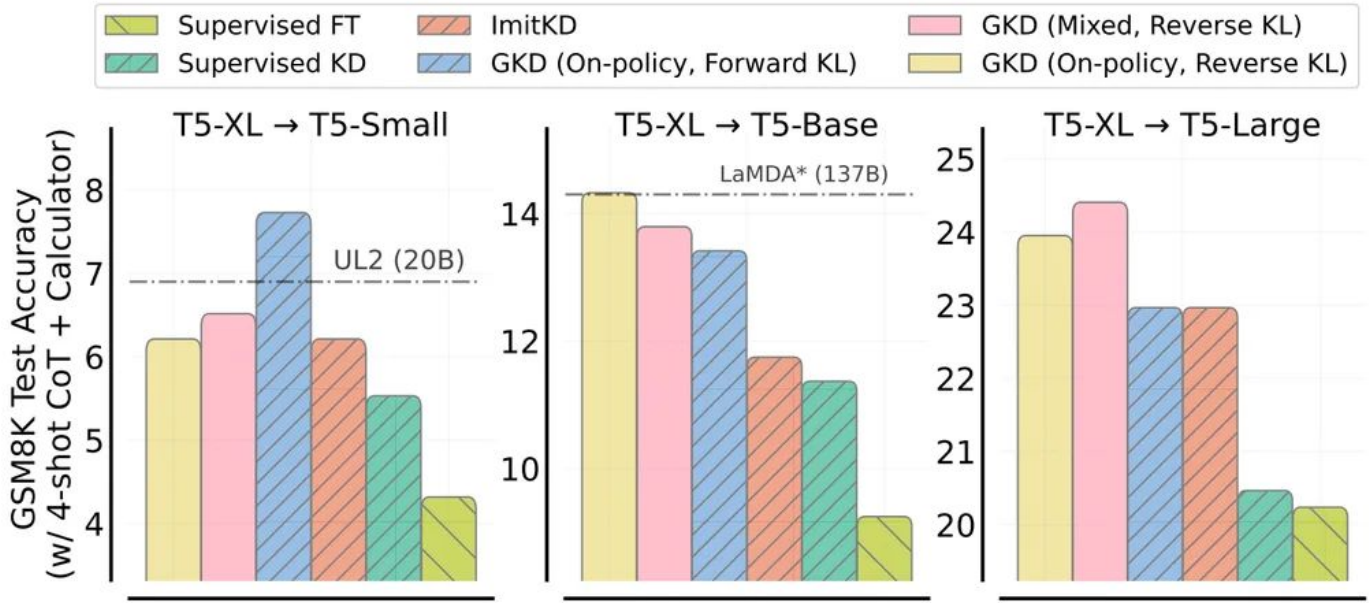


Results: Machine Translation (WMT en-de)





Results: Mathematical Reasoning (GSM8K) with CoT



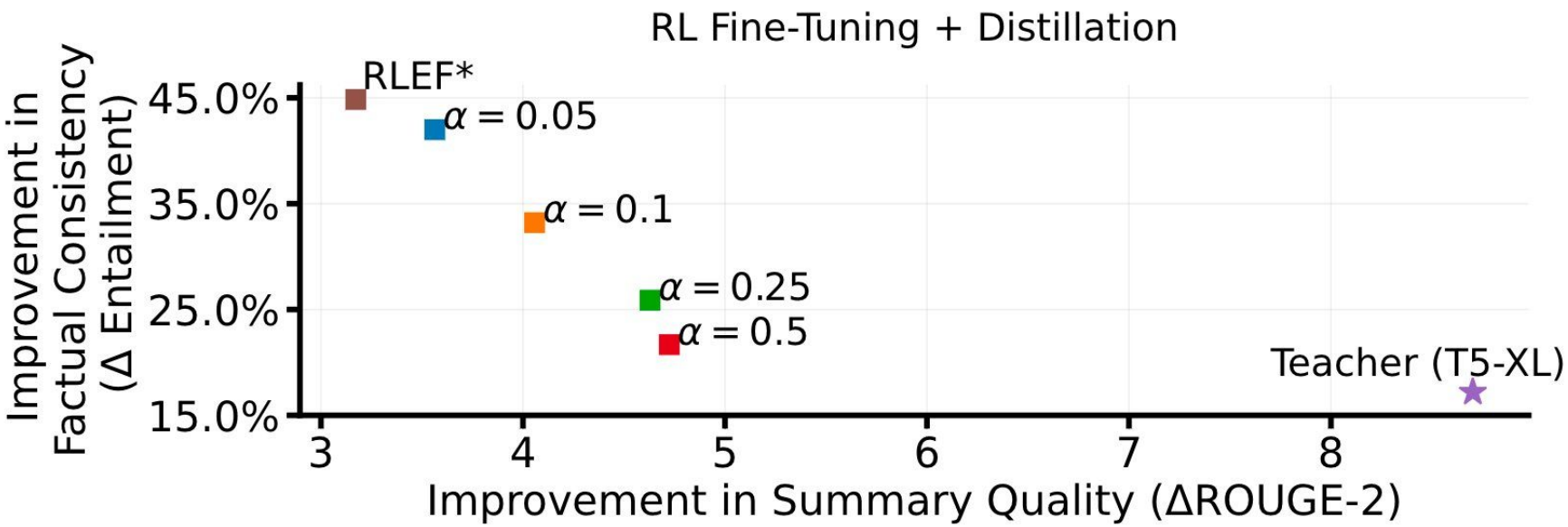
GKD + RLHF: Cherry on Cake

Rishabh Agarwal
@agarwl_

[1/3] Here's a simple but powerful idea to improve RLHF / RLAIF by combining it with knowledge distillation: Simply regularize the LLM policy (student) to a more capable teacher model instead of the base student model for the KL regularization term.

$$(1-\alpha)\underbrace{E_{y\sim p_S^\theta(\cdot|x)}\left[r(y)\right]}_{\text{RL objective}} - \alpha\underbrace{\mathbb{E}_{y\sim p_S(\cdot|x)}\left[\mathcal{D}(p_T\|p_S^\theta)\right.}_{\text{Generalized On-Policy Disti}}$$

GKD + RLHF: Cherry on Cake



α controls the strength of the distillation.

A future direction I am excited about ..

Diffusion Models can be viewed as autoregressive models

- Can we do better distillation of such models? Maybe!

Feel free to reach out if interested in collaborating!