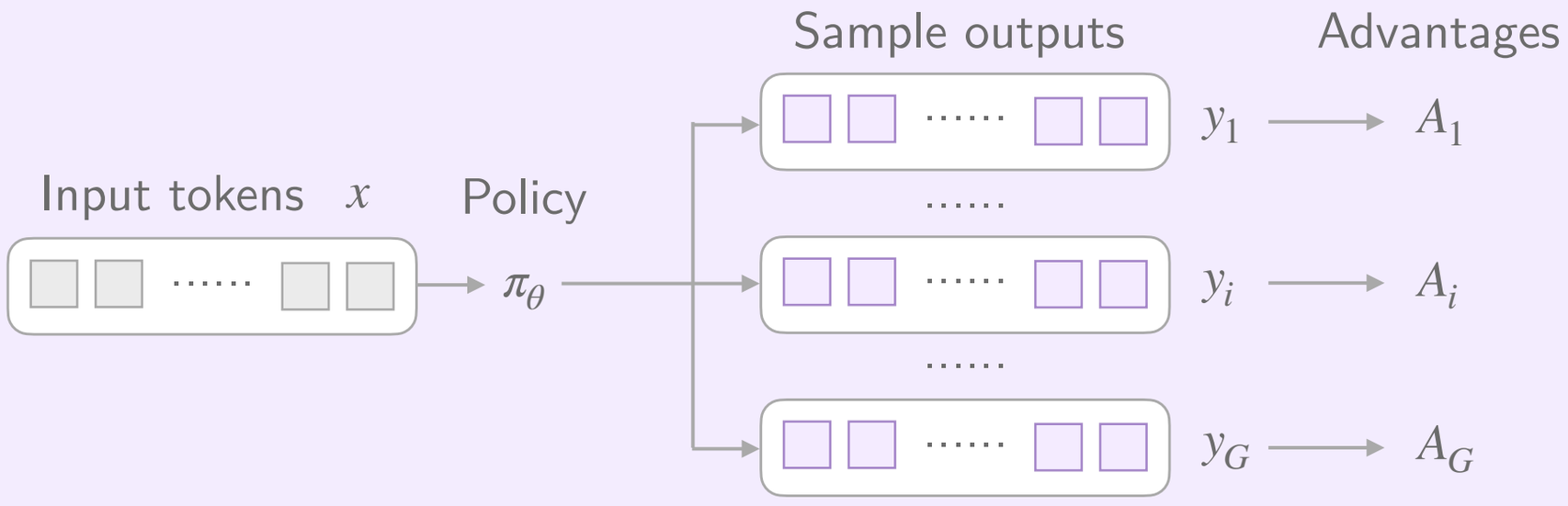
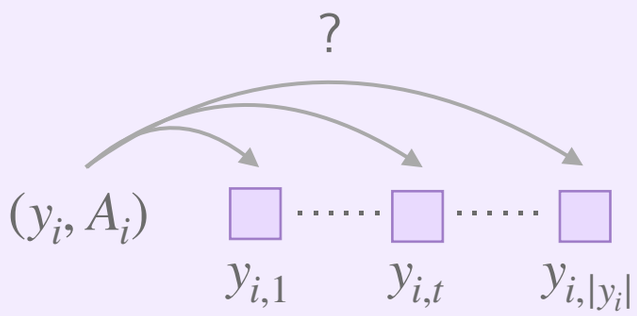
GRPO

GRPO sampling



GRPO optimization

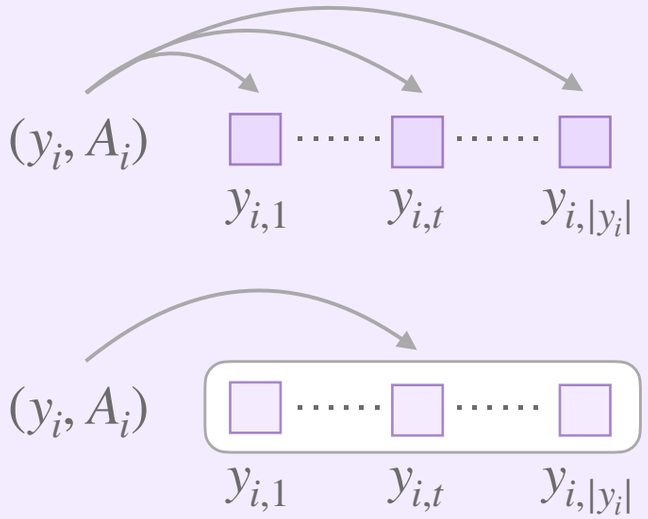


$$w_{i,t} = \frac{\pi_{\theta'}(y_{i,t} \mid x, y_{i,<t})}{\pi_\theta(y_{i,t} \mid x, y_{i,<t})} \qquad w_{i,t} \cdot A_i$$

$$J_{\text{GRPO}}(\theta) = \mathbb{E}\left(\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|y_i|}\sum_{t=1}^{|y_i|} w_{i,t} \cdot A_i\right)$$

GRPO → GSPO

GSPO optimization



$$w_{i,t} = \frac{\pi_{\theta'}(y_{i,t} \mid x, y_{i,<t})}{\pi_{\theta}(y_{i,t} \mid x, y_{i,<t})} \qquad w_{i,t} \cdot A_i \qquad \text{GRPO}$$

$$w_i = \left(\frac{\pi_{\theta'}(y_i \mid x)}{\pi_{\theta}(y_i \mid x)}\right)^{\frac{1}{|y_i|}} \qquad w_{i,t} \cdot A_i \qquad \text{GSPO}$$

$$J_{\text{GRPO}}(\theta) = \mathbb{E}\left(\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|y_i|}\sum_{t=1}^{|y_i|} w_{i,t} \cdot A_i\right)$$

$$J_{\text{GSPO}}(\theta) = \mathbb{E}\left(\frac{1}{G}\sum_{i=1}^{G} w_i \cdot A_i\right)$$

Related works

$$J_{\text{GRPO}}(\theta) = \mathbb{E}\left(\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|y_i|}\sum_{t=1}^{|y_i|} w_{i,t}\cdot A_i\right)$$

$$J_{\text{GSPO}}(\theta) = \mathbb{E}\left(\frac{1}{G}\sum_{i=1}^{G} w_i\cdot A_i\right)$$ Qwen Team (25.07)

$$J_{\text{GMPO}}(\theta) = \mathbb{E}\left(\frac{1}{G}\sum_{i=1}^{G}\prod_{t=1}^{|y_i|}[w_{i,t}\cdot A_i]^{\frac{1}{|y_i|}}\right)$$ Geometric-mean Policy Optimization (25.07), token-level

$$J_{\text{P-GSPO}}(\theta) = \mathbb{E}\left(\frac{1}{G}\sum_{i=1}^{G} w'_i\cdot A_i\right) \qquad w_i = \left(\frac{\pi_{\theta'}(y_i|x)}{\pi_\theta(y_i|x)}\right)^{\frac{1}{|y_i|}} \qquad w'_i = \left(\frac{\pi_{\theta'}(y_i|x)}{\pi_\theta(y_i|x)}\right)^{\frac{1}{|y_i|^{\alpha}}}$$ Parameterized GSPO (25.10), sequence-level

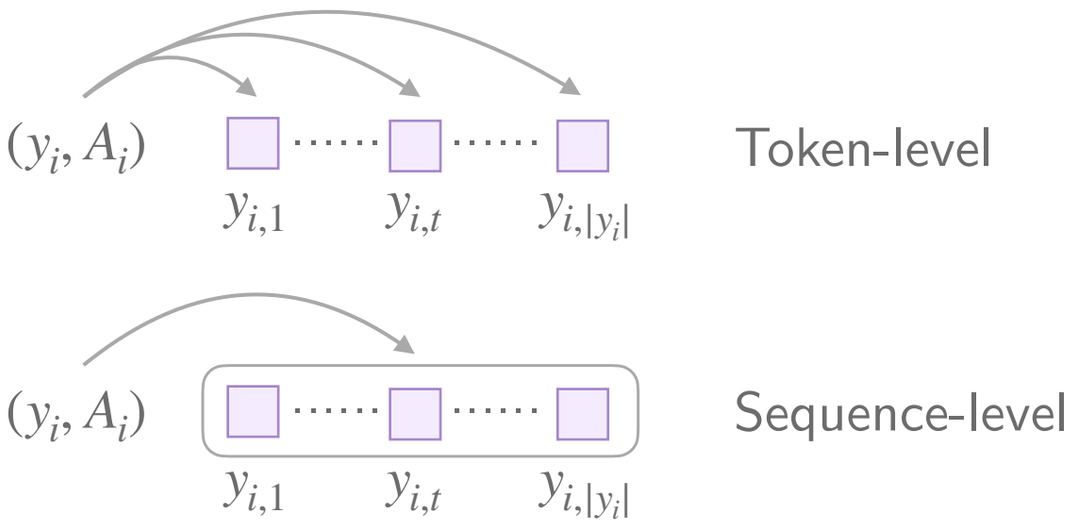
$$J_{\text{SAPO}}(\theta) = \begin{cases} \mathbb{E}\left(\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|y_i|}\sum_{t=1}^{|y_i|} w_{i,t}\cdot A_i\right), \text{smooth sequence} \\ \mathbb{E}\left(\frac{1}{G}\sum_{i=1}^{G} w_i\cdot A_i\right), \text{nonsmooth sequence} \end{cases}$$

Soft Adaptive Policy Optimization
Qwen Team (25.11), mixed-level

Discussions

- Token-level 的问题：理论上针对每一个 token 计算 importance score 时，不满足独立同分布的条件，因此逐 token 的 importance score 无法进行严格有效的分布校正。但它提供了细粒度控制。
- Sequence-level 的问题：理论上正确，序列才是自然的采样单位，满足独立同分布的条件，但是粒度过粗。一个序列中所有 token 共用一个标量权重，导致对长度不敏感，并丢失序列内部的 credit assignment 能力。
- 目标：我们既需要稳定且正确的 credit assignment，也需要更高效率地进行优化。

Related works



Discussions

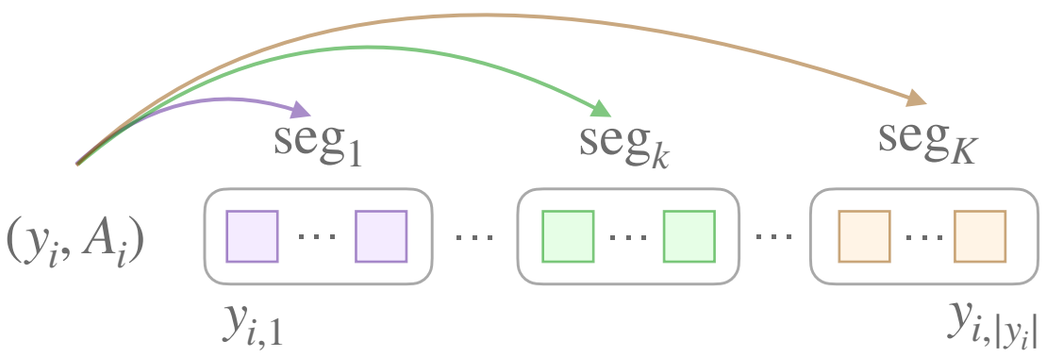
- Text token 特性：相邻 token 依赖相对较弱。（证明）
- Speech token 特性：相邻 token 依赖相对较强。（证明）
- 总结：speech token 的重要特征是连续的 token 之间具有相比 text token 更强的关联性。我们的目标是找到一个匹配 speech token 特性的，可以动态作出分段的优化策略。
- 目标：
  - 更匹配 reward 粒度。 对于语音，reward 通常是在整句层面评估的。但句子的不同部分对 reward 的贡献可能不同。一个出现发音错误的片段，与一个发音正确但韵律异常的片段，应当有不同的 importance score 分配。分段能做到这一点，而 sequence level 无法区分。
  - 缓解长度归一化问题。 GSPO 对整条序列做归一化，导致对长度不敏感。分段级归一化只在每个片段内部进行，随后再对片段求平均。这样既能保留对片段数量（间接对应长度）的敏感性，又能在片段内部维持归一化稳定性。
  - 总体上具有 GSPO 稳定的优势和更精细的 credit assignment 优势，optimization 效率更高。

Demo designs

$$w_{i,t} = \frac{\pi_{\theta'}(y_{i,t} \mid x, y_{i,<t})}{\pi_{\theta}(y_{i,t} \mid x, y_{i,<t})} \qquad w_{i,t} \cdot A_i \qquad \text{GRPO}$$

$$w_i = \left(\frac{\pi_{\theta'}(y_i \mid x)}{\pi_{\theta}(y_i \mid x)}\right)^{\frac{1}{|y_i|}} \qquad w_{i,t} \cdot A_i \qquad \text{GSPO}$$

$$w_{i,k} = \left(\sum_{t \in \text{seg}_k} \frac{\pi_{\theta'}(y_{i,t} \mid x, y_{i,<t})}{\pi_{\theta}(y_{i,t} \mid x, y_{i,<t})}\right)^{\frac{1}{|\text{seg}_k|}} \qquad w_{i,t} \cdot A_i = w_{i,k} \cdot A_i, i \in \text{seg}_k \qquad \text{Our}$$



Demo designs

- (Baseline) Fixed window segmentation: 按固定窗口大小 W 切分成 ⌈T/W⌉ 个不重叠的 segment。每个 segment 内的所有 token 共享同一个 importance score
- Log-probability segmentation: 计算相邻 tokens 之间的 logprob 差异，如果大于一个 threshold 就判定为 segmentation
- TokenVal segmentation：计算相邻 tokens 之间的 token ID 差异，如果大于一个 threshold 就判定为 segmentation
- ... (brainstorm)

Demo

- 在 HKUSTAudio/Llasa-1B (https://huggingface.co/HKUSTAudio/Llasa-1B) 上进行了快速尝试
- (Check WandB)
- 初步效果：加了分割之后确实会比 GRPO 有效；直接应用 GSPO 反而效果会变差一些。
- 潜在问题：
  - HuggingFace 上有人在 SkyRL 中实现了 SAPO，但在小规模 dense model（Qwen3-4B）上发现 SAPO 跑不过 DAPO baseline. 一种可能是在小规模 dense model 上，token 之间出现很高方差的情况不多，在这种情况下 sequence level 的 optimization 的优势不明显。
  - 目前的分割方法比较基础，没有针对 speech 更加特化的分割方式