RL's Role in Hierarchical Structures: RL can be beneficial if the goal is to introduce a decision-making process that goes beyond standard supervised learning. For instance, RL could help the model make a sequence of decisions that involve dependencies among labels (e.g., group depends on supergroup). However, this approach is typically more useful in complex environments where choices impact later predictions.

Policy Gradient in Current Implementation: The current setup uses a standard CrossEntropyLoss for both the supervised and RL policy logits, essentially treating RL as a classification problem without leveraging actual policy gradients (e.g., rewards from actions) that are common in RL tasks. This approach doesn’t capture the unique strengths of RL, like optimizing for long-term rewards.

Where RL Could be Useful:

Decision Dependencies: If there are complex dependencies between hierarchy levels, RL could allow the model to "choose" groups based on prior decisions, which might yield some advantage.
Sequential Decision-Making: In scenarios where predicting each hierarchy level is sequential and depends on the prior decisions, a proper policy gradient approach could allow the model to learn decision paths that improve accuracy over time.