OOD Generalization Gap: Dev (In-Domain) vs Eval (Out-of-Domain)