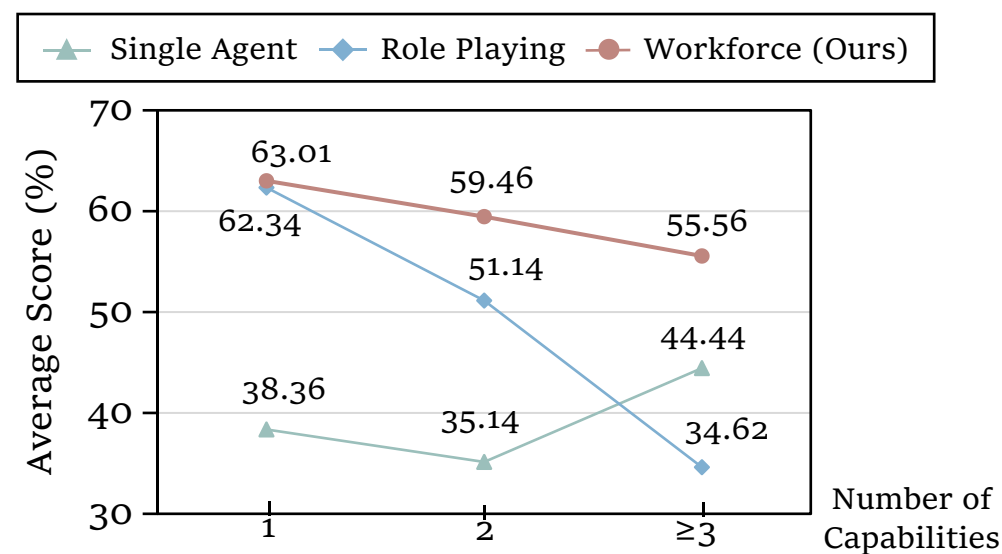
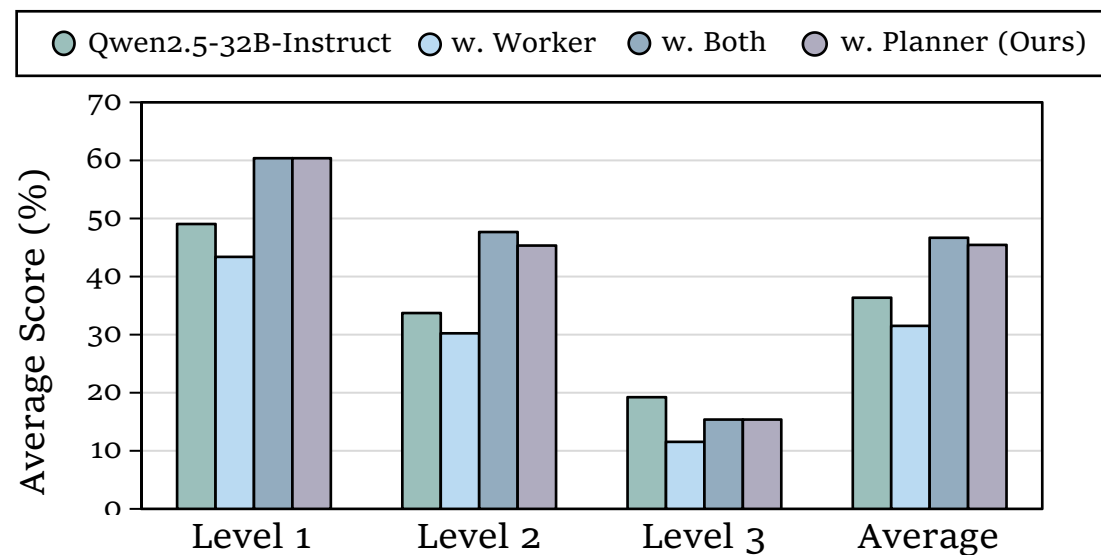


(a) Failure Modes on the GAIA Benchmark



(b) Performance across different capability requirements



(c) Performance comparason between different training configuration