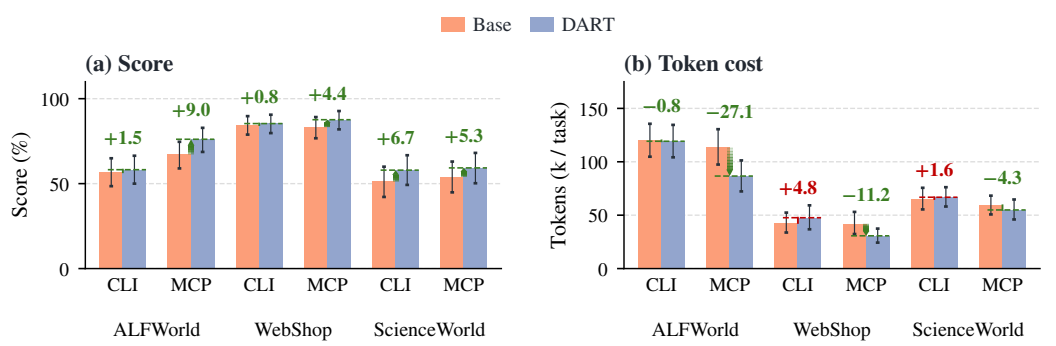
Base
DART
(a) Score
Score (%)
100
50
0
+1.5
+9.0
+0.8
+4.4
+6.7
+5.3
CLI
MCP
CLI
MCP
CLI
MCP
ALFWorld
WebShop
ScienceWorld
(b) Token cost
Tokens (k / task)
150
100
50
0
−0.8
−27.1
+4.8
−11.2
+1.6
−4.3
CLI
MCP
CLI
MCP
CLI
MCP
ALFWorld
WebShop
ScienceWorld