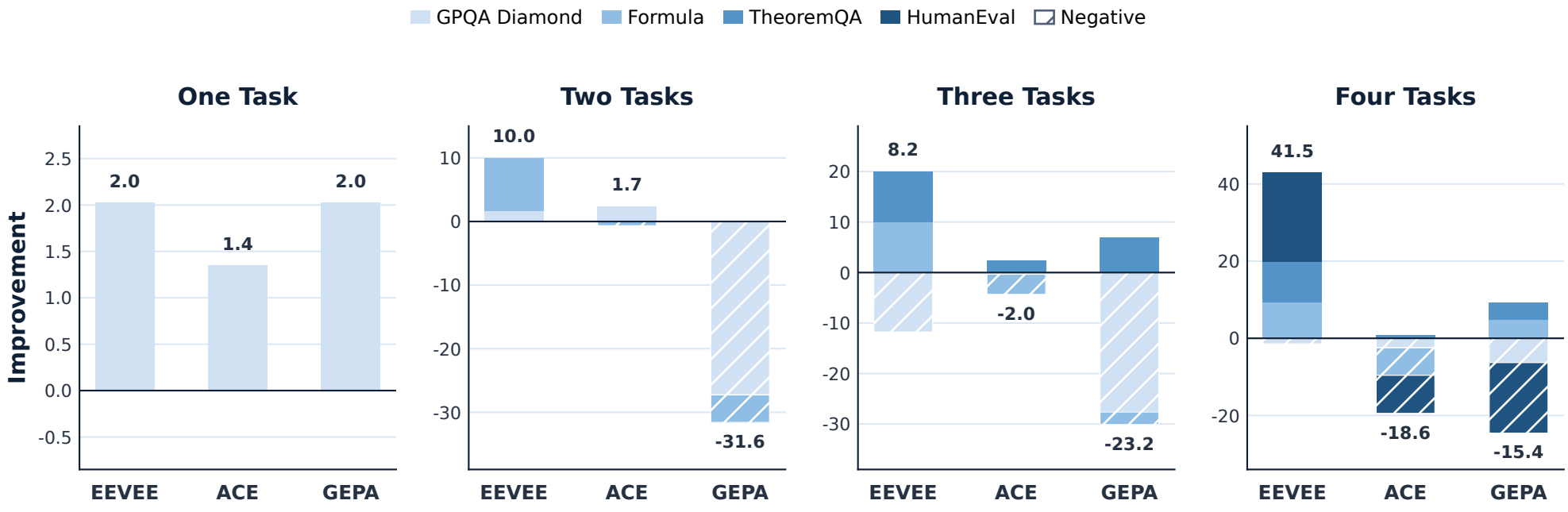
GPQA Diamond
Formula
TheoremQA
HumanEval
Negative
One Task
Two Tasks
Three Tasks
Four Tasks
Improvement
2.5
2.0
1.5
1.0
0.5
0.0
-0.5
2.0
1.4
2.0
10
0
-10
-20
-30
10.0
1.7
-31.6
20
10
0
-10
-20
-30
8.2
-2.0
-23.2
40
20
0
-20
41.5
-18.6
-15.4
EEVEE
ACE
GEPA