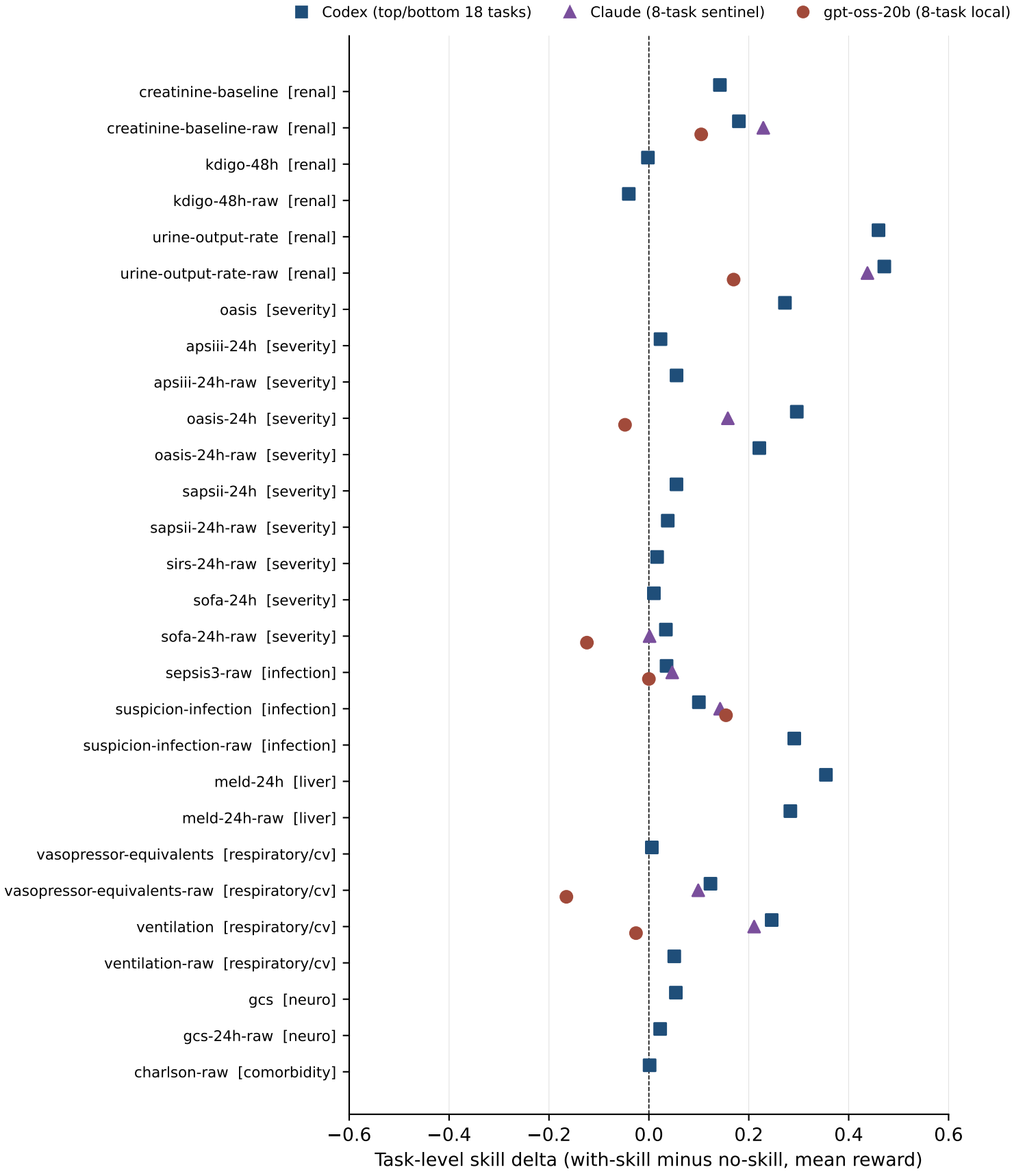
Codex (top/bottom 18 tasks)
Claude (8-task sentinel)
gpt-oss-20b (8-task local)
creatinine-baseline [renal]
creatinine-baseline-raw [renal]
kdigo-48h [renal]
kdigo-48h-raw [renal]
urine-output-rate [renal]
urine-output-rate-raw [renal]
oasis [severity]
apsiii-24h [severity]
apsiii-24h-raw [severity]
oasis-24h [severity]
oasis-24h-raw [severity]
sapsii-24h [severity]
sapsii-24h-raw [severity]
sirs-24h-raw [severity]
sofa-24h [severity]
sofa-24h-raw [severity]
sepsis3-raw [infection]
suspicion-infection [infection]
suspicion-infection-raw [infection]
meld-24h [liver]
meld-24h-raw [liver]
vasopressor-equivalents [respiratory/cv]
vasopressor-equivalents-raw [respiratory/cv]
ventilation [respiratory/cv]
ventilation-raw [respiratory/cv]
gcs [neuro]
gcs-24h-raw [neuro]
charlson-raw [comorbidity]
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6
Task-level skill delta (with-skill minus no-skill, mean reward)