(a) Fixed-15 · Weak set · classify_objects_by_language, layout 4: wooden toys → left, chocolate → middle, peppers → right



three baskets; peppers, chocolate bars and wooden toys on the table; the policy starts



the pepper is held but heads for the left basket; GPT-6 carries it toward the centre



the right arm holds the red pepper above the red basket; the left hand grasps the elephant



white chocolate over the middle basket; two targets refused, the robot holds for 25 steps



chocolates and peppers sorted, the wooden toys still on the table (fail, Score 40)

Fixed-15
fail: step limit 1100
Score 40
42 calls
20 takeovers
15 continues
7 refused, 1 hold
policy 510 steps
GPT-6 565 steps
hold 25 steps



(b) Fixed-15 · Weak set · organize_table, layout 1: mouse → pad, keyboard → frame, figurine → stand, clock → drawer, open it



the left arm approaches the figurine; clock, mouse and keyboard are still out of place



the clock slipped during the lift; GPT-6 takes over 4 steps to close the gripper fully



four items placed, arms idle at home; GPT-6 moves the left gripper to the drawer handle

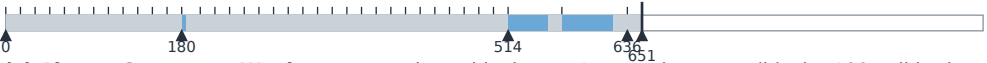


second handle reach 7 cm short, the drawer stays closed; GPT-6 lets the arm withdraw



give_up after two failed drawer reaches; the robot idles to step 1000 (Score 75)

Fixed-15
fail: give_up at 651
Score 75
38 calls
3 takeovers
34 continues
policy 555 steps
GPT-6 96 steps
idle 651–1000



(c) Always-Segment · Weak set · organize_table, layout 1: same layout as (b); the 100-call budget is spent at step 935



GPT-6 alone from the first call: the right gripper turns down toward the mouse; four items out of place



mouse on the pad (call 25) and clock on the drawer (call 52) after repeated grasps; five reaches refused



the figurine slips from the left gripper over the keyboard; two regrasps (calls 62–70) fail



calls 84–97 are one-step nudges at the keyboard's left edge; it stays outside the frame



call 100 returns both arms home at step 935; nothing more executes before the step limit (Score 25)

Always-Segment
fail: budget spent at 935
Score 25
100 calls (budget)
95 move_eef segments
5 refused
no policy steps
idle 935–1000

