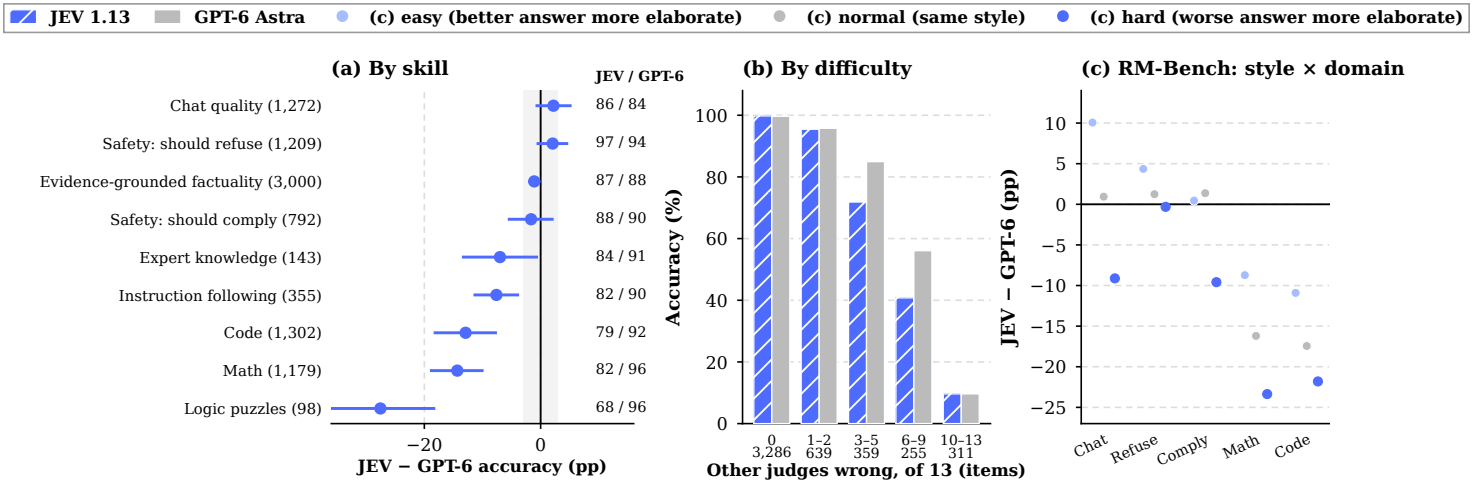

JEV 1.13
GPT-6 Astra
(c) easy (better answer more elaborate)
(c) normal (same style)
(c) hard (worse answer more elaborate)
(a) By skill
JEV / GPT-6
Chat quality (1,272)
86 / 84
Safety: should refuse (1,209)
97 / 94
Evidence-grounded factuality (3,000)
87 / 88
Safety: should comply (792)
88 / 90
Expert knowledge (143)
84 / 91
Instruction following (355)
82 / 90
Code (1,302)
79 / 92
Math (1,179)
82 / 96
Logic puzzles (98)
68 / 96
−20
0
JEV − GPT-6 accuracy (pp)
(b) By difficulty
Accuracy (%)
100
80
60
40
20
0
0
3,286
1–2
639
3–5
359
6–9
255
10–13
311
Other judges wrong, of 13 (items)
(c) RM-Bench: style × domain
JEV − GPT-6 (pp)
10
5
0
−5
−10
−15
−20
−25
Chat
Refuse
Comply
Math
Code