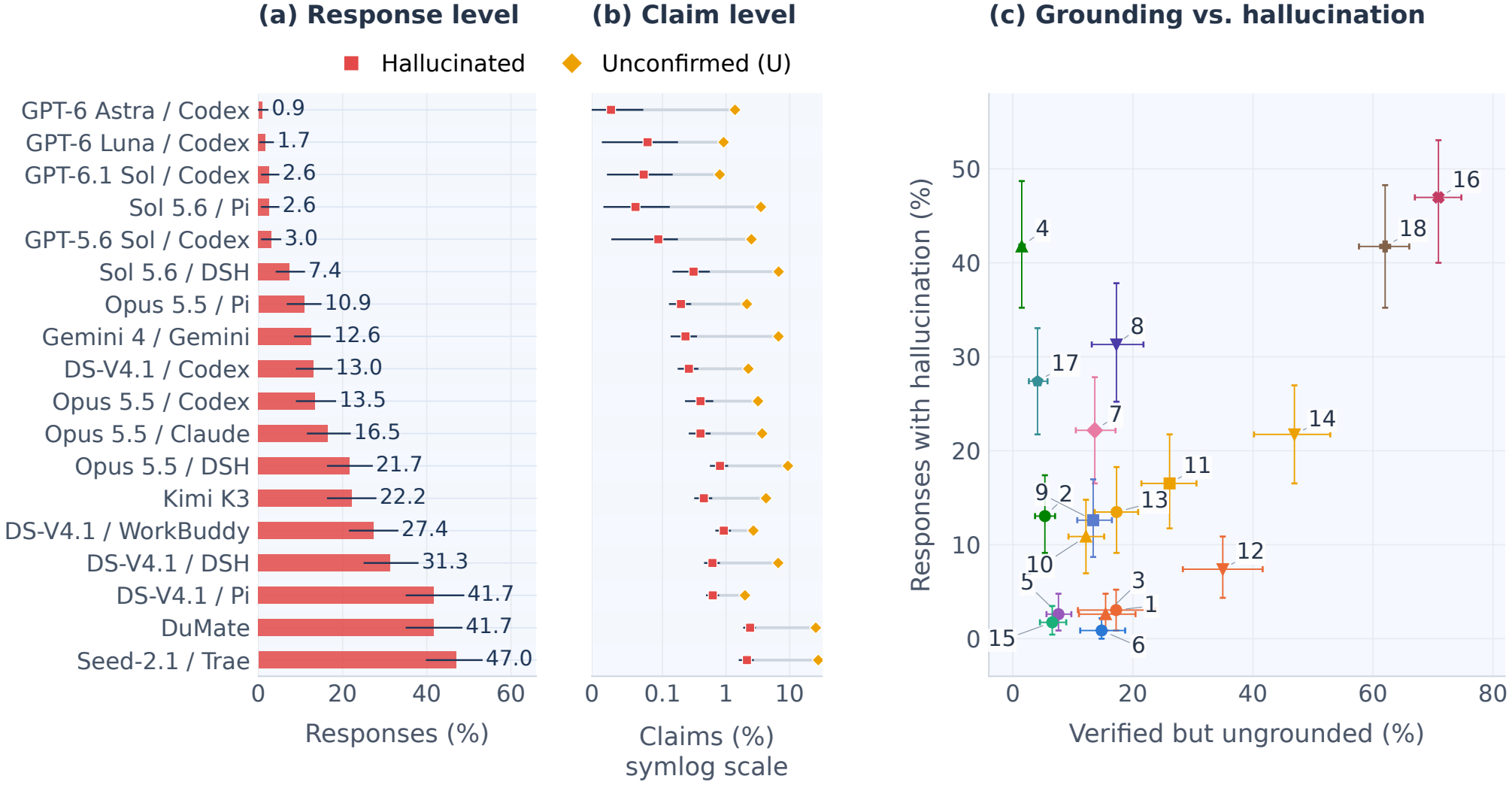

(a) Response level
(b) Claim level
(c) Grounding vs. hallucination
Hallucinated
Unconfirmed (U)
GPT-6 Astra / Codex
GPT-6 Luna / Codex
GPT-6.1 Sol / Codex
Sol 5.6 / Pi
GPT-5.6 Sol / Codex
Sol 5.6 / DSH
Opus 5.5 / Pi
Gemini 4 / Gemini
DS-V4.1 / Codex
Opus 5.5 / Codex
Opus 5.5 / Claude
Opus 5.5 / DSH
Kimi K3
DS-V4.1 / WorkBuddy
DS-V4.1 / DSH
DS-V4.1 / Pi
DuMate
Seed-2.1 / Trae
0.9
1.7
2.6
2.6
3.0
7.4
10.9
12.6
13.0
13.5
16.5
21.7
22.2
27.4
31.3
41.7
41.7
47.0
0
20
40
60
Responses (%)
0
0.1
1
10
Claims (%)
symlog scale
Responses with hallucination (%)
Verified but ungrounded (%)
0
10
20
30
40
50
0
20
40
60
80
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18