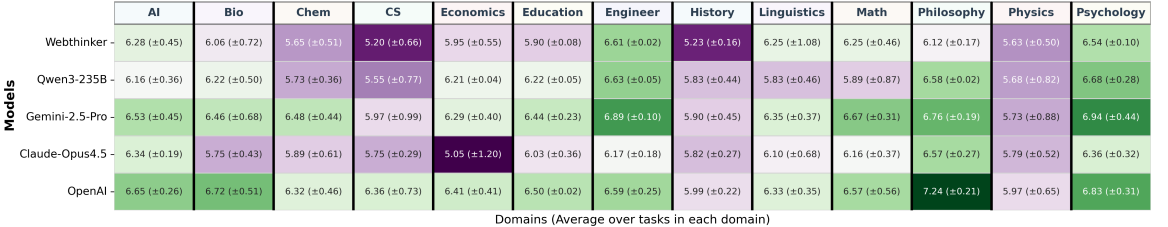

Request Fulfillment
Analytical Soundness
Structural Coherence
Format & Style
Information Integrity
Information Sufficiency
Ethics & Compliance
Models
Webthinker 4.71 3.59 4.81 5.01 4.54 6.85 5.47 5.36 5.47 7.99 7.23 6.82 7.04 3.57 3.26 7.80 6.97 7.75 9.02 7.28 6.22 2.00 6.50 8.42 5.81
Qwen3-235B 4.36 3.50 4.66 4.98 4.56 5.45 4.63 4.22 5.49 7.20 6.85 6.97 6.80 5.95 5.01 5.43 6.41 8.86 9.03 8.54 8.18 4.34 6.70 8.23 5.94
Gemini-2.5-Pro 5.04 3.82 5.01 6.06 5.02 6.91 6.05 6.32 6.29 7.63 7.20 7.24 7.28 5.16 4.49 4.02 6.50 9.18 8.47 9.08 7.04 4.76 6.19 8.42 5.96
Claude-Opus4.5 4.87 3.61 5.04 5.51 5.10 6.05 5.21 6.51 5.91 7.30 7.18 6.84 7.13 4.39 4.08 6.08 6.11 8.81 8.32 7.38 3.64 3.08 6.31 8.48 6.38
OpenAI 5.21 3.82 5.21 5.57 5.04 7.24 5.98 6.57 5.77 7.87 7.21 7.09 7.35 6.04 5.45 8.27 6.97 7.29 7.34 9.92 7.94 1.94 6.93 8.72 6.36
completeness
scope
helpfulness
quantification
reasoning
intro
body
conclusion
section
report_fmt
writing
paragraph
readability
claim_fact
cite_support
ref_support
ref_quality
ref_diversity
evid_cover
info_amount
cite_amount
ref_amount
sensitive
safety
balance
Score
10
9
8
7
6
5
4
3
2


AI
Bio
Chem
CS
Economics
Education
Engineer
History
Linguistics
Math
Philosophy
Physics
Psychology
Models
Webthinker 6.28 (±0.45) 6.06 (±0.72) 5.65 (±0.51) 5.20 (±0.66) 5.95 (±0.55) 5.90 (±0.08) 6.61 (±0.02) 5.23 (±0.16) 6.25 (±1.08) 6.25 (±0.46) 6.12 (±0.17) 5.63 (±0.50) 6.54 (±0.10)
Qwen3-235B 6.16 (±0.36) 6.22 (±0.50) 5.73 (±0.36) 5.55 (±0.77) 6.21 (±0.04) 6.22 (±0.05) 6.63 (±0.05) 5.83 (±0.44) 5.83 (±0.46) 5.89 (±0.87) 6.58 (±0.02) 5.68 (±0.82) 6.68 (±0.28)
Gemini-2.5-Pro 6.53 (±0.45) 6.46 (±0.68) 6.48 (±0.44) 5.97 (±0.99) 6.29 (±0.40) 6.44 (±0.23) 6.89 (±0.10) 5.90 (±0.45) 6.35 (±0.37) 6.67 (±0.31) 6.76 (±0.19) 5.73 (±0.88) 6.94 (±0.44)
Claude-Opus4.5 6.34 (±0.19) 5.75 (±0.43) 5.89 (±0.61) 5.75 (±0.29) 5.05 (±1.20) 6.03 (±0.36) 6.17 (±0.18) 5.82 (±0.27) 6.10 (±0.68) 6.16 (±0.37) 6.57 (±0.27) 5.79 (±0.52) 6.36 (±0.32)
OpenAI 6.65 (±0.26) 6.72 (±0.51) 6.32 (±0.46) 6.36 (±0.73) 6.41 (±0.41) 6.50 (±0.02) 6.59 (±0.25) 5.99 (±0.22) 6.33 (±0.35) 6.57 (±0.56) 7.24 (±0.21) 5.97 (±0.65) 6.83 (±0.31)
Average Score
7.0
6.5
6.0
5.5
Domains (Average over tasks in each domain)