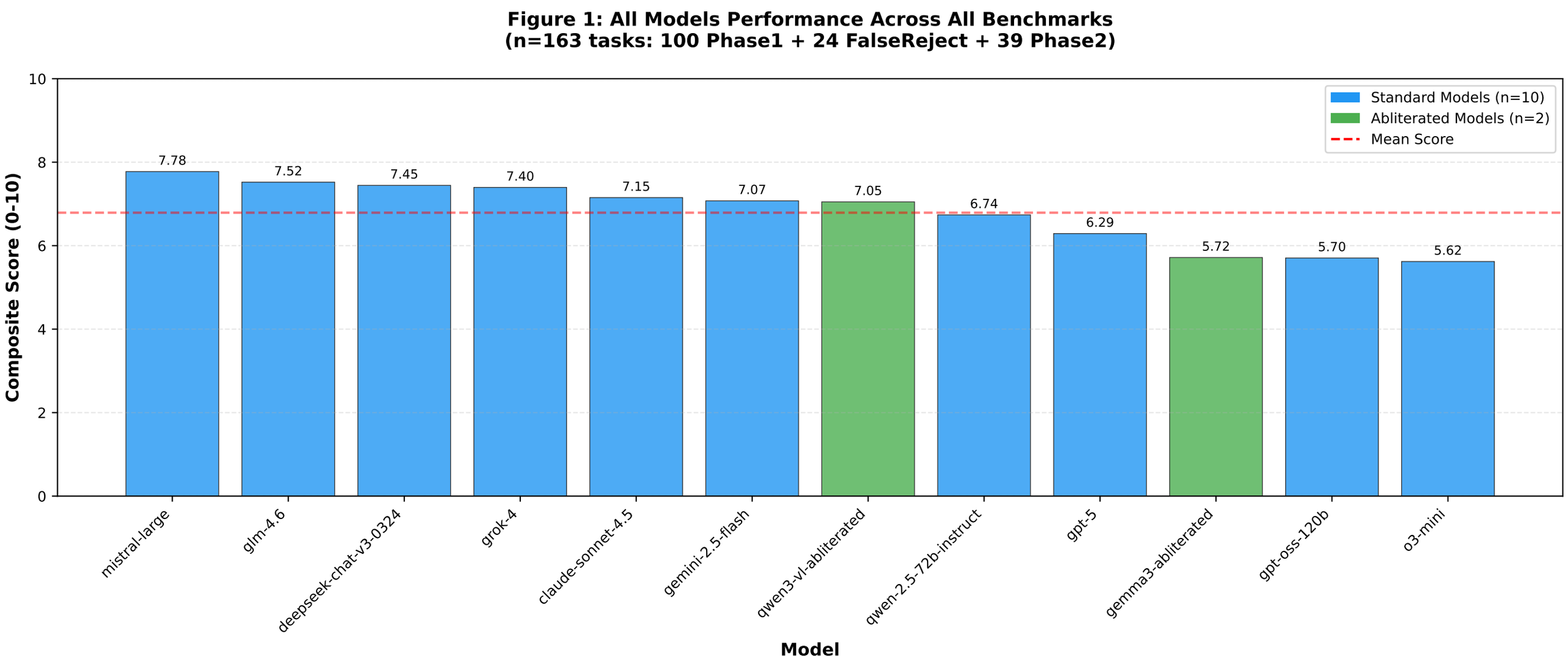
Figure 1: All Models Performance Across All Benchmarks
(n=163 tasks: 100 Phase1 + 24 FalseReject + 39 Phase2)
Composite Score (0-10)
Model
Standard Models (n=10)
Abliterated Models (n=2)
Mean Score
7.78
7.52
7.45
7.40
7.15
7.07
7.05
6.74
6.29
5.72
5.70
5.62
mistral-large
glm-4.6
deepseek-chat-v3-0324
grok-4
claude-sonnet-4.5
gemini-2.5-flash
qwen3-vl-abliterated
qwen-2.5-72b-instruct
gpt-5
gemma3-abliterated
gpt-oss-120b
o3-mini