Cross-Model Three-Tier Open-Ended Evaluation (N = 1,032)

Legend:
- Level A (Correct)
- Level B (Partial)
- Level C (Incorrect)

GPT-4o-mini: Level A (Correct) 24.5%, Level B (Partial) 21.5%, Level C (Incorrect) 54.0%; Strict: 24.5%, Lenient: 46.0%

GPT-5-mini: Level A (Correct) 41.8%, Level B (Partial) 22.3%, Level C (Incorrect) 35.9%; Strict: 41.8%, Lenient: 64.1%