# Judge Score Statistics Summary (with_hil)

| Model | Count | Mean | Median | Std | High Conf.% | Low Conf.% |
|---|---|---|---|---|---|---|
| GPT-4o-mini | 9 | 0.978 | 1.000 | 0.063 | 100.0% | 0.0% |
| Claude 3.7 Sonnet | 95 | 0.857 | 0.800 | 0.101 | 85.3% | 0.0% |
| DeepSeek V3 0324 | 15 | 0.967 | 1.000 | 0.047 | 100.0% | 0.0% |