Hybrid Model Evaluation - GSM8K

Accuracy (%)

- Base: 31.7% (38/120)
- Hybrid: 63.3% (76/120), +31.7% (64% of Gap)
- Thinking: 80.8% (97/120), 49.2% Gap