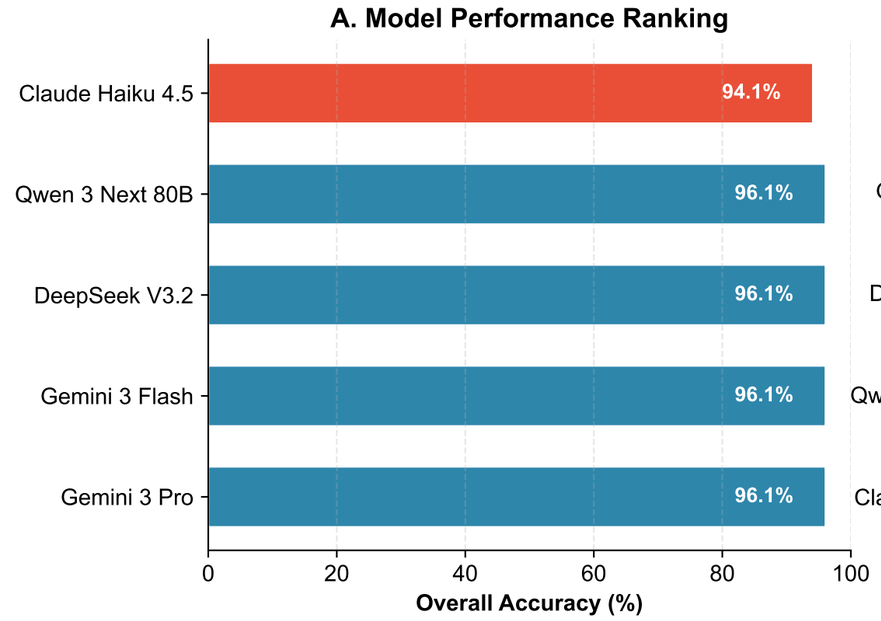
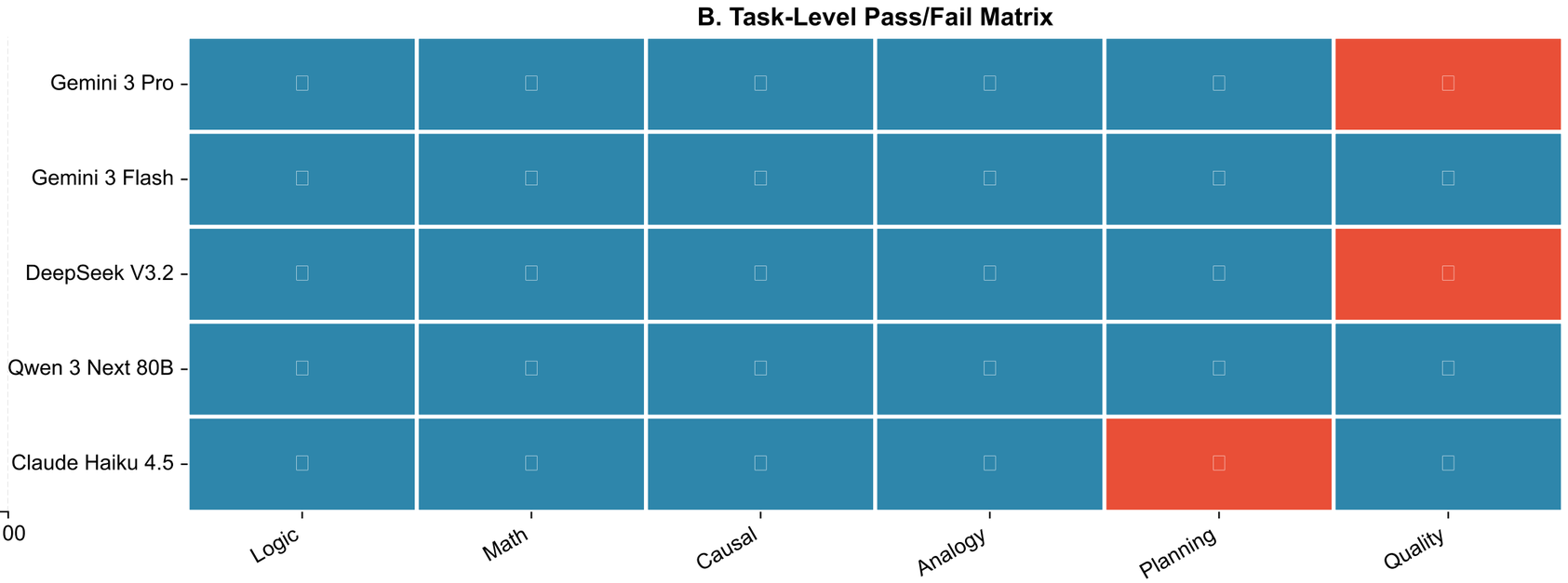
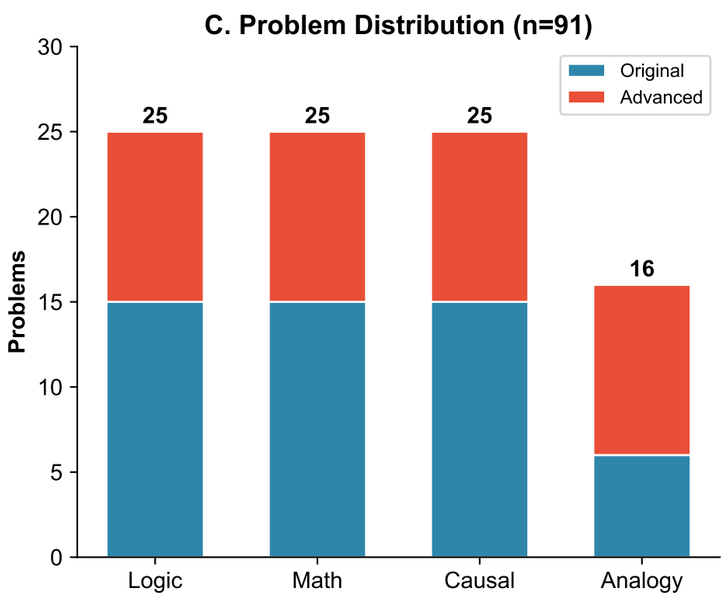
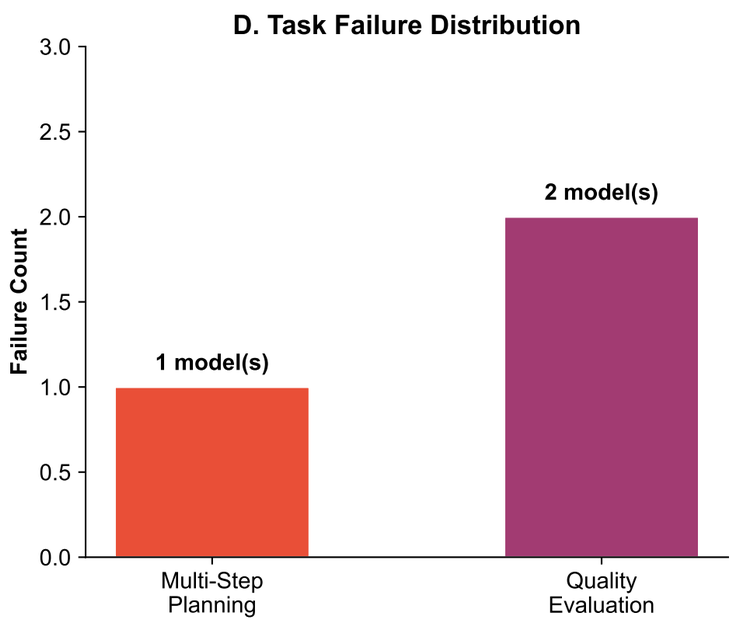
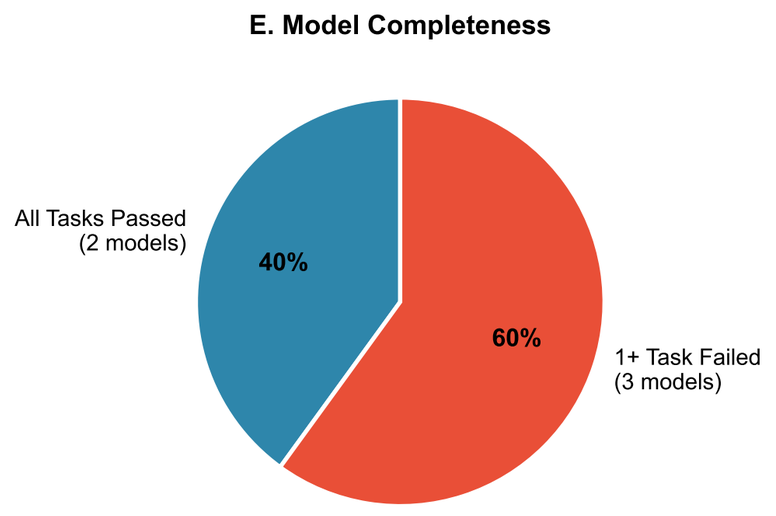
CORE-Bench: Comprehensive LLM Reasoning Evaluation Dashboard











KEY FINDINGS FROM CORE-BENCH EVALUATION

 OVERALL PERFORMANCE
• 4 of 5 models achieved 96.08% accuracy
• Claude Haiku 4.5 scored 94.12% (lowest)
• All models passed core reasoning tasks

 TOP PERFORMERS
• Gemini 3 Flash: Perfect on all 6 tasks
• Qwen 3 Next 80B: Perfect on all 6 tasks
• Both demonstrated robust reasoning

⚠ CHALLENGE AREAS
• Quality Evaluation: 2 failures (40%)
• Multi-Step Planning: 1 failure (20%)
• Core reasoning tasks: 0 failures

 BENCHMARK STATISTICS
• Total: 91 reasoning problems
• 4 categories tested
• 6 evaluation tasks + 1 comprehensive

 ADVANCED PROBLEMS
• 40 advanced problems added (44% of total)
• Simpson's Paradox, Modal Logic, etc.
• Designed to challenge frontier models

 RECOMMENDATIONS
• Re-run with expanded 91-problem set
• Focus testing on Quality Evaluation
• Add per-problem accuracy tracking