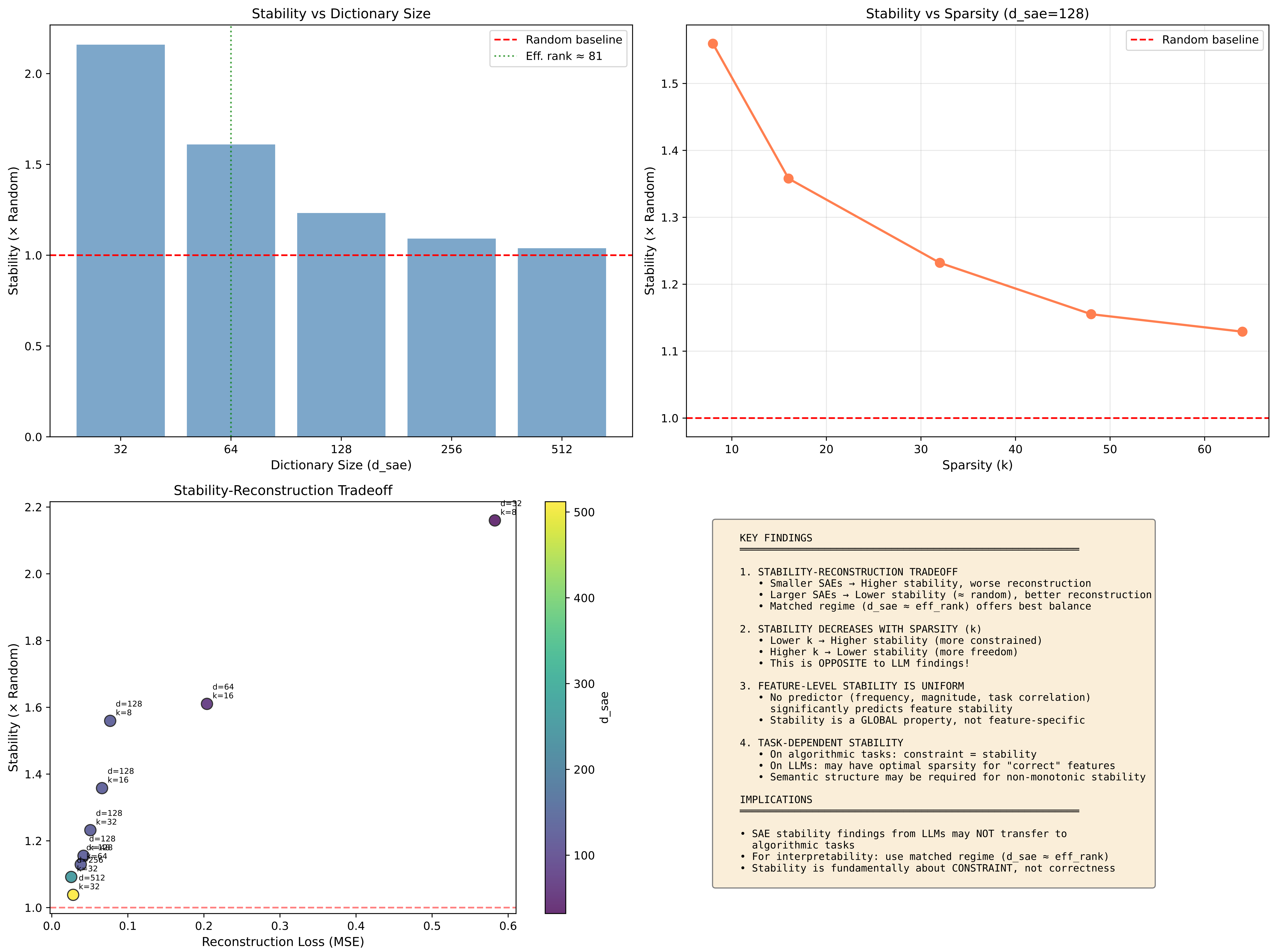

```
KEY FINDINGS
═══════════════════════════════════════════════════

1. STABILITY-RECONSTRUCTION TRADEOFF
   • Smaller SAEs → Higher stability, worse reconstruction
   • Larger SAEs → Lower stability (≈ random), better reconstruction
   • Matched regime (d_sae ≈ eff_rank) offers best balance

2. STABILITY DECREASES WITH SPARSITY (k)
   • Lower k → Higher stability (more constrained)
   • Higher k → Lower stability (more freedom)
   • This is OPPOSITE to LLM findings!

3. FEATURE-LEVEL STABILITY IS UNIFORM
   • No predictor (frequency, magnitude, task correlation)
     significantly predicts feature stability
   • Stability is a GLOBAL property, not feature-specific

4. TASK-DEPENDENT STABILITY
   • On algorithmic tasks: constraint = stability
   • On LLMs: may have optimal sparsity for "correct" features
   • Semantic structure may be required for non-monotonic stability

IMPLICATIONS
═══════════════════════════════════════════════════

• SAE stability findings from LLMs may NOT transfer to
  algorithmic tasks
• For interpretability: use matched regime (d_sae ≈ eff_rank)
• Stability is fundamentally about CONSTRAINT, not correctness
```