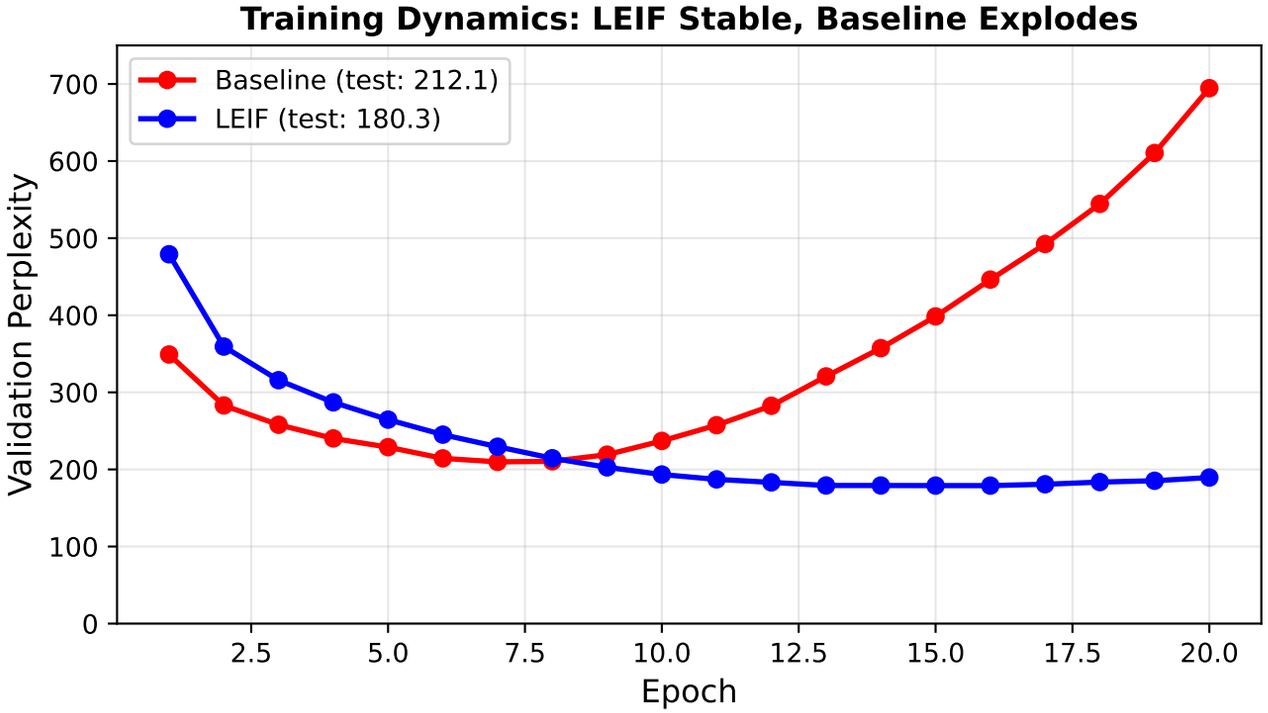
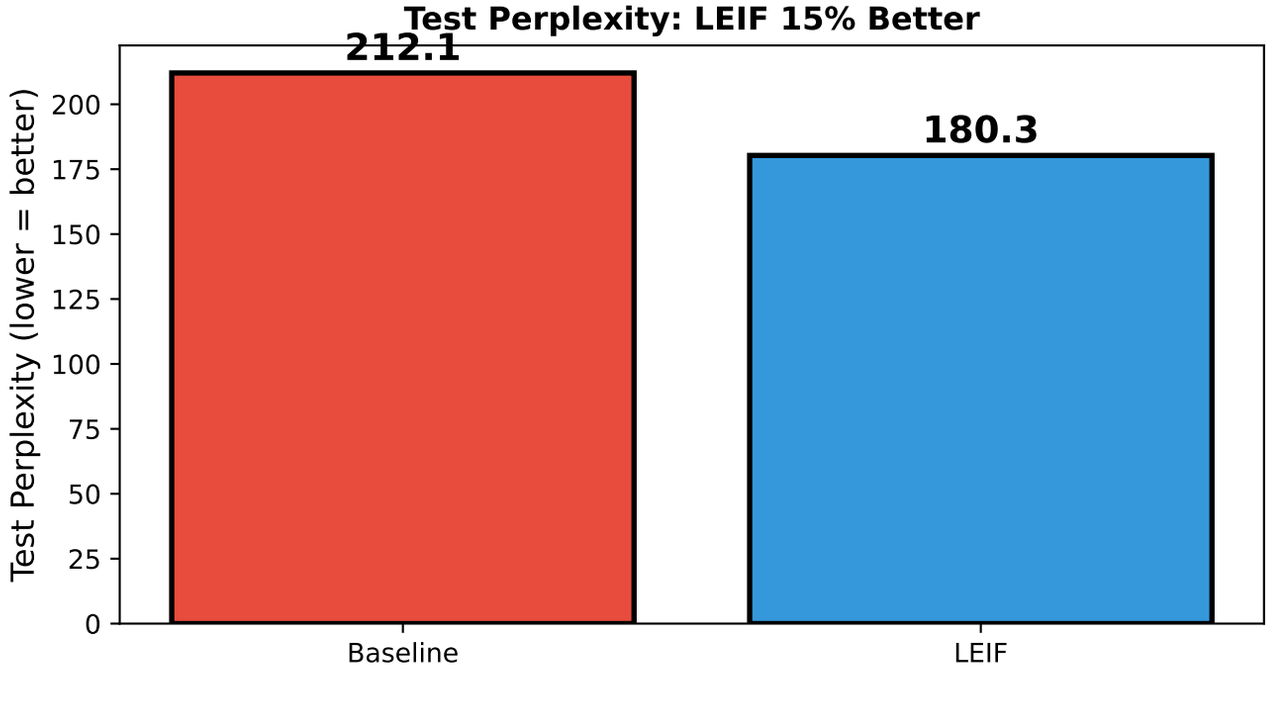
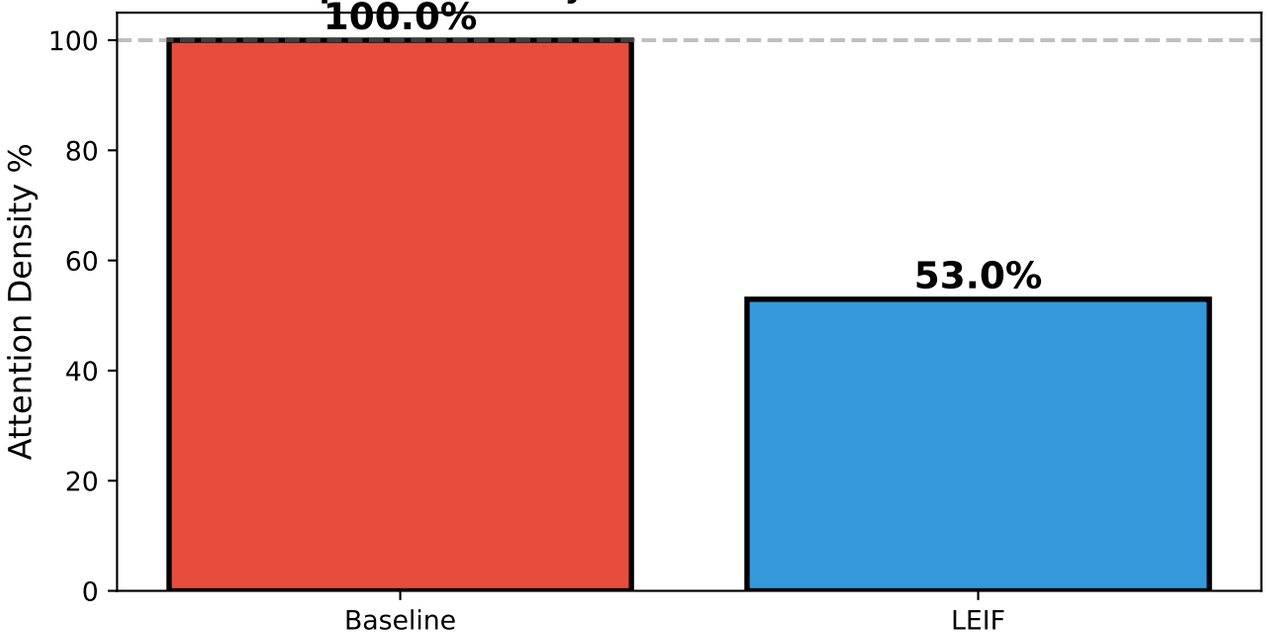
LEIF vs Baseline: Real Ubuntu IRC Multi-Party Benchmark
5000 conversations, 4-6 agents per window, 512 seq_len








```
BENCHMARK SUMMARY
═══════════════════════════════════════════

Dataset: Ubuntu IRC Dialogue Corpus
         5,000 multi-party conversations
         4-6 agents per window (multiparty sampling)
         512 token sequence length

RESULTS:

┌──────────────────┬──────────┬──────────┐
│ Metric           │ Baseline │ LEIF     │
├──────────────────┼──────────┼──────────┤
│ Test Perplexity  │ 212.1    │ 180.3    │
│ Best Val PPL     │ 209.8    │ 179.0    │
│ Final Val PPL    │ 694.6    │ 189.5    │
│ Attention %      │ 100%     │ 52.9%    │
│ Parameters       │ 17.9M    │ 15.0M    │
└──────────────────┴──────────┴──────────┘

KEY FINDINGS:
• LEIF achieves 15% lower perplexity
• LEIF uses 47% less attention compute
• Baseline catastrophically overfits
• LEIF training remains stable

CONCLUSION: Sparse relational attention
            outperforms dense attention
            on multi-party dialogue.
```