

TODO: Token limit problems

Priority: High

Due: October 19

Problem:

- long docs exceed LM context
- Llama-2-7b has 4K token limit
- Current chunking: 500 Chrs, 100 overlap
- Summaries still too long

Actions:

- Benchmark chunk effectiveness
 - Test overlap percentages
 - Measure token distribution
 - Consider dynamic chunking
 - Explore compression techniques
- Bob is leading the optimization effort
- Target: Oct 25
- Effort: 4-6 hours