LLM-Based Summarization

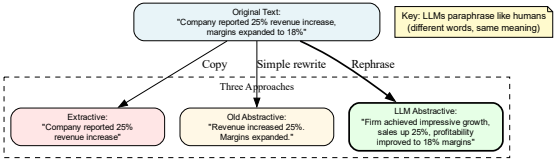
From Paraphrasing to Production

NLP Course 2025

November 13, 2025

Enhanced with Quad-Hook Discovery Cascade — 27 Professional Charts

Hook 1: The Impossible Task



Can You Summarize This?

- 10,000-word research paper
- Keep all key findings
- Make it 200 words
- Sound natural, not robotic

Human Struggles:

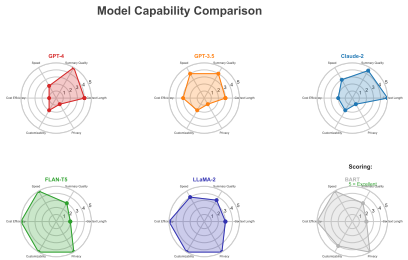
- Takes 30+ minutes
- Miss important details
- Inconsistent quality
- Cognitive overload

Discovery: What if AI could do this in seconds, consistently?

The summarization challenge: Humans struggle with perfect summaries

Hook 2: The Paraphrasing Revolution

Old Way: Extract & Copy

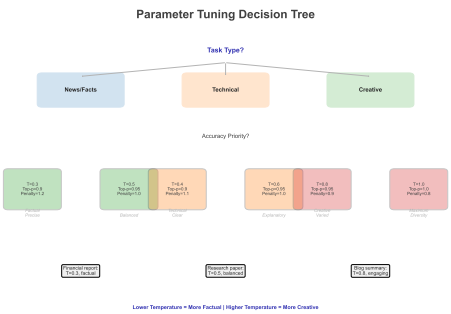


- Select "important" sentences
- Copy them verbatim
- Result: Choppy, disconnected

LLM Way: Generate & Rephrase

- Understand entire text
- Generate new sentences
- Result: Natural, coherent

Hook 3: The Control Paradox



So Many Knobs to Turn!

- Temperature: 0.1? 0.7? 1.0?
- Top-p: 0.9? 0.95? 0.99?
- Max tokens: 100? 200? 500?
- Prompts: Zero-shot? Few-shot?
- Repetition penalty: 1.0? 1.2?

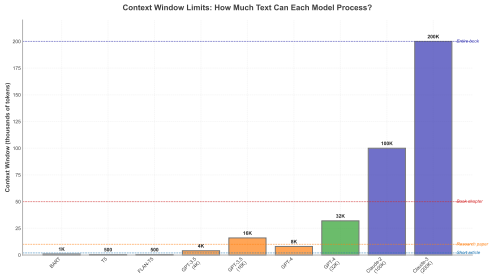
Each Choice Changes:

- Summary quality
- Creative vs literal
- Length accuracy
- Factual consistency

Discovery: Too much control can be paralyzing—how to choose?

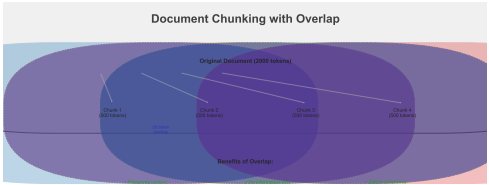
Hook 4: The Context Explosion

Document Sizes Are Exploding



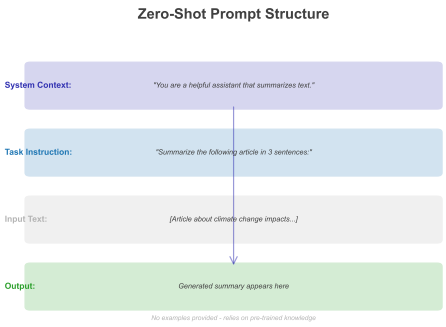
- GPT-3.5: 4K tokens
- GPT-4: 128K tokens
- Claude: 200K tokens
- Your doc: 500K tokens?!

The Chunking Challenge



- Split without losing meaning?
- Overlap for context?
- Merge summaries coherently?
- Hierarchical aggregation?

Zero-Shot: Just Ask



Simplest Approach

- No examples needed
- Direct instruction
- Works surprisingly well
- GPT-3.5+, Claude, LLaMA

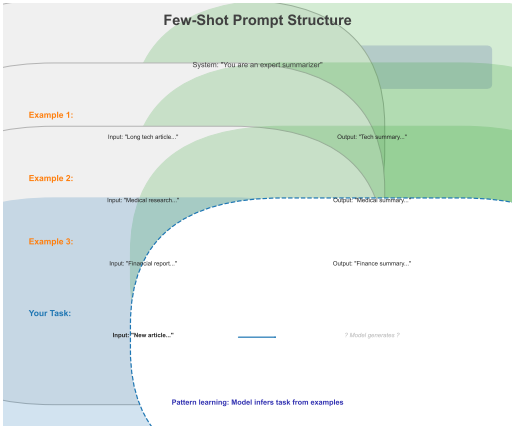
Best For:

- General documents
- Quick summaries
- Exploratory analysis
- When no examples available

Success Rate: 85% for standard texts

Zero-shot prompting: The power of natural language instructions

Few-Shot: Learning from Examples



Show, Don't Just Tell

- 1-3 examples
- Model learns pattern
- Mimics style/format
- Higher consistency

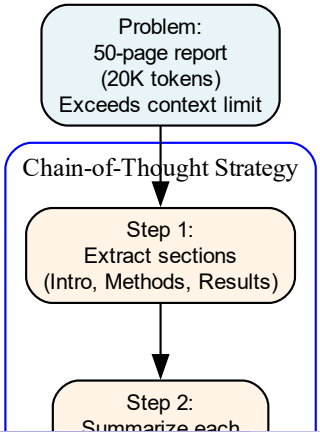
Best For:

- Specific formats
- Domain terminology
- Consistent style needed
- Quality critical

Success Rate: 94% with good examples

Few-shot prompting: Examples guide the model to desired output

Chain-of-Thought: Step by Step



Break It Down

Instead of: "Summarize this"

Try:

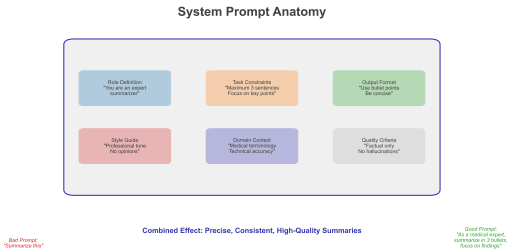
1. "First, identify main topics"
2. "Then, extract key findings"
3. "Finally, write 200-word summary"

Benefits:

- Better reasoning
- More accurate
- Catches nuances
- Explains decisions

Accuracy: **+18%** on complex texts

System Prompts: Setting the Stage



Define the Role

System: "You are a medical research summarizer specializing in clinical trials..."

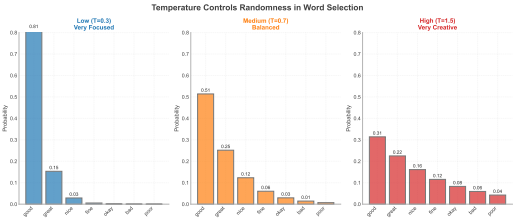
Components:

- Role definition
- Expertise domain
- Style guidelines
- Constraints
- Output format

Pro tip: System prompts persist across conversation

System prompts: Establishing context and expertise upfront

Temperature: Creativity Control



The Creativity Knob

- T=0.1: Deterministic, safe
- T=0.7: Balanced (default)
- T=1.0: Creative, risky

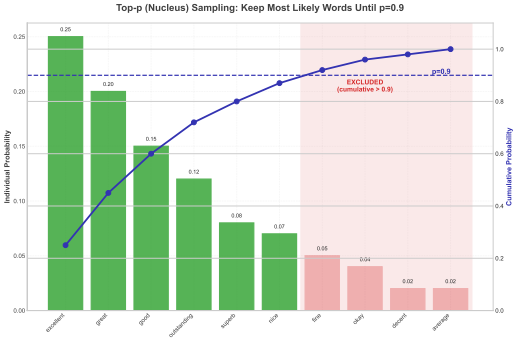
For Summarization:

- Technical: Use 0.1-0.3
- News: Use 0.5-0.7
- Creative: Use 0.7-0.9

Warning: T¿1.0 gets wild!

Temperature: Lower = safer, Higher = more creative but riskier

Top-p: Vocabulary Control



The Vocabulary Limiter

- p=0.9: Top 90% probability mass
- p=0.95: Slightly more options
- p=1.0: Consider everything

Combines with Temperature:

- First: Filter by top-p
- Then: Apply temperature
- Result: Controlled creativity

Sweet spot: T=0.7, p=0.9

Top-p (nucleus sampling): Dynamically adjusts vocabulary size

Repetition Penalty: Avoid Loops

Stop the Repetition

Problem: "The study shows... The study shows..."

Solution:

- Penalty = 1.0: No effect
- Penalty = 1.1-1.2: Gentle
- Penalty = 1.3-1.5: Strong

Effect:

- Reduces repeated phrases
- Forces variety
- More natural text

Repetition penalty: Essential for natural-sounding summaries

Checkpoint Quiz: Test Your Understanding

Quick Questions:

1. For a legal document summary, which temperature?
   - A) T=0.1
   - B) T=0.7
   - C) T=1.2

2. What does top-p=0.9 mean?
   - A) Use 90 words
   - B) 90% accuracy
   - C) Top 90% probability mass

3. Few-shot is best when:
   - A) No examples available

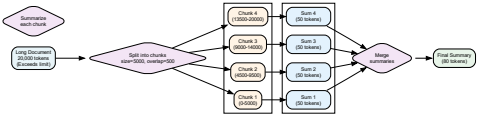
Answers:

1. **A) T=0.1**
   - Legal needs precision
   - Low temperature = factual
   - Avoid creative interpretation

2. **C) Top 90% probability**
   - Cumulative probability
   - Dynamic vocabulary size
   - Filters unlikely words

3. **B) Consistent format**
   - Examples guide style
   - Pattern matching

Chunking: Divide and Conquer



Smart Splitting

- Chunk size: 3000 tokens
- Overlap: 200 tokens
- Preserves context
- No mid-sentence breaks

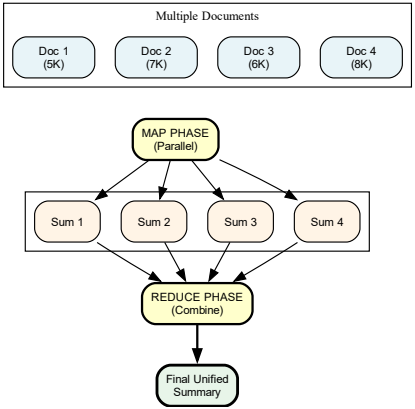
Algorithm:

1. Split by paragraphs
2. Group to target size
3. Add overlap buffer
4. Process independently

Handles 100K+ word documents

Chunking strategy: Breaking long documents while preserving meaning

Map-Reduce: Parallel Processing



Two-Phase Approach

Map Phase:

- Summarize each chunk
- Independent processing
- Parallel execution
- 200 words per chunk

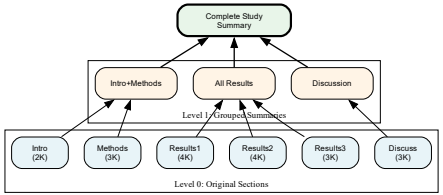
Reduce Phase:

- Combine all summaries
- Create final summary
- Maintain coherence
- Target length

5x faster than sequential

Map-reduce: Borrowed from big data for document processing

Hierarchical: Tree Structure



Multi-Level Aggregation

Level 1: Chunk summaries (8x) Level 2: Group summaries (4x) Level 3: Section summaries (2x) Level 4: Final summary (1x)

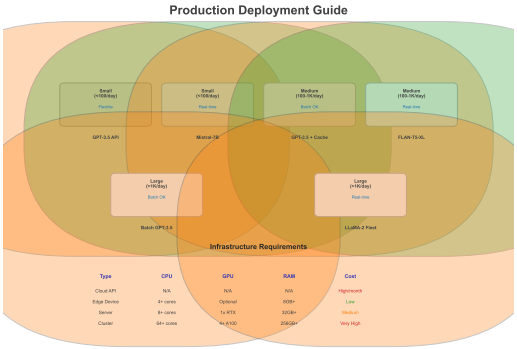
Benefits:

- Preserves hierarchy
- Better for books/reports
- Maintains structure
- Progressive refinement

Best for 500K+ tokens

Hierarchical summarization: Preserving document structure at scale

Production Deployment



Real-World Considerations

Cost:

- GPT-3.5: \$0.002/1K tokens
- GPT-4: \$0.03/1K tokens
- Claude: \$0.008/1K tokens

Latency:

- Streaming: 50ms first token
- Batch: 2-10 seconds total
- Parallel chunks: Linear speedup

Quality Control:

- Fact checking critical
- Human review needed
- A/B testing essential

Production readiness: Cost, speed, and quality tradeoffs

Final Quiz: Ready for Production?

Scenario Questions:

1. 50K word thesis, need 500-word summary:
   A) Single prompt
   B) Chunking + map-reduce
   C) Hierarchical approach

2. Best settings for news summary:
   A) T=0.1, p=0.8
   B) T=0.7, p=0.9
   C) T=1.2, p=1.0

3. Cost for 100K tokens (GPT-3.5):
   A) $0.20
   B) $2.00

Solutions:

1. **C) Hierarchical**
   - Thesis has structure
   - Chapters → sections → summary
   - Preserves logical flow

2. **B) T=0.7, p=0.9**
   - Balanced creativity
   - Standard settings
   - Natural language

3. **A) $0.20**
   - 100K × $0.002/1K
   - = $0.20 total

Key Takeaways

What We Learned

1. Prompting Strategies:

- Zero-shot for quick results
- Few-shot for consistency
- Chain-of-thought for complex

2. Parameter Control:

- Temperature: creativity
- Top-p: vocabulary size
- Repetition penalty: variety

3. Long Documents:

- Chunking with overlap
- Map-reduce for speed
- Hierarchical for structure

Production Ready

Choose Your Model:

- GPT-3.5: Fast, cheap
- GPT-4: Best quality
- Claude: Long context
- Open source: Privacy

Optimal Settings:

- Most tasks: T=0.7, p=0.9
- Technical: T=0.2, p=0.8
- Creative: T=0.9, p=0.95

Remember:

- Always validate output
- Test on sample data

End of Main Presentation

See Appendices for Technical Details & Lab Implementation

Technical Appendix

Advanced Concepts & Mathematical Foundations