Token Usage Optimization Checklist

1. Audit Token Usage:
   - Track input/output token counts
   - Identify top-consuming workloads
   - Separate by model tier
2. Choose the Right Model Tier:
   - High reasoning -> GPT-5 or Gemini Pro
   - Medium complexity -> GPT-5 mini, Gemini Flash
   - Simple tasks -> GPT-5 nano, Flash-Lite, Mistral Medium
3. Compress Prompts:
   - Shorten instructions
   - Remove repeated boilerplate
   - Use abbreviations or IDs
4. Reduce Output Length:
   - Set max_tokens limits
   - Use concise style instructions
   - Return raw JSON or CSV when possible
5. Use Prompt Caching:
   - Cache repeated context where supported
   - Simulate caching in GPT-5 by storing reusable context
6. Batch & Chunk:
   - Combine small requests into one
   - Chunk large documents and summarize first
7. Pre-Process Before the LLM:
   - Use regex, keyword extraction, or filtering
   - Pass only relevant data to the model
8. Monitor & Iterate:
   - Set usage alerts
   - A/B test prompt variations
   - Review workloads quarterly