Context Windows, API Costs, and Token Limits in LLMs

LLM Context Window

- Context window = max tokens model can process at once
- Includes all conversation history, not just latest message
- Each generated token adds to the input for next step
- Limits how much background/model memory is available
- Important for multi-shot prompting, RAG, inference techniques
- Larger context windows allow more references, e.g., ticket prices, full documents

Model Examples and Context Window Sizes

- GPT-5: 400,000 token context window
- Claude: 200,000 tokens
- GPT OSS (open source): ~130,000 tokens
- Gemini 2.5 Flash: 1,000,000 tokens (can fit nearly all of Shakespeare)
- Larger windows enable more complex prompts and memory

API Costs

- Chat products: free and paid tiers, monthly subscriptions
- API usage: pay per use, regardless of subscription
- Costs based on input + output tokens
- Input tokens = full conversation history, any inserted memory, RAG content
- Output tokens include model's reasoning (even if not shown)
- Example: GPT-5 input $1.25/million tokens, output $10/million tokens
- Small API calls (few tokens) are very cheap
- GPT-5 Nano: input $0.05/million tokens, output $0.40/million tokens

Cost Considerations

- Costs can add up with large context or many users
- Caching can reduce costs for repeated inputs (automatic with GPT, varies with Claude)
- Important to estimate per-user costs for scaling
- Individual experimentation is low-cost unless building large systems

Tools and Resources

- Vellum AI LLM Leaderboard: compare models, see context windows and API costs
- Useful for understanding model capabilities and pricing

Upcoming Topics

- Will cover leaderboards, prompting techniques, streaming results
- Next sessions: hands-on with OpenAI API, chat completions, business solution implementation
- Will learn about OneShot prompting, markdown/JSON streaming, illusion of memory

General Takeaways

- Understanding context window and API costs is key for LLM use
- Larger context windows = more powerful, but higher cost
- Small-scale use is affordable; scaling requires careful cost management
- Ready for practical, hands-on work in next session