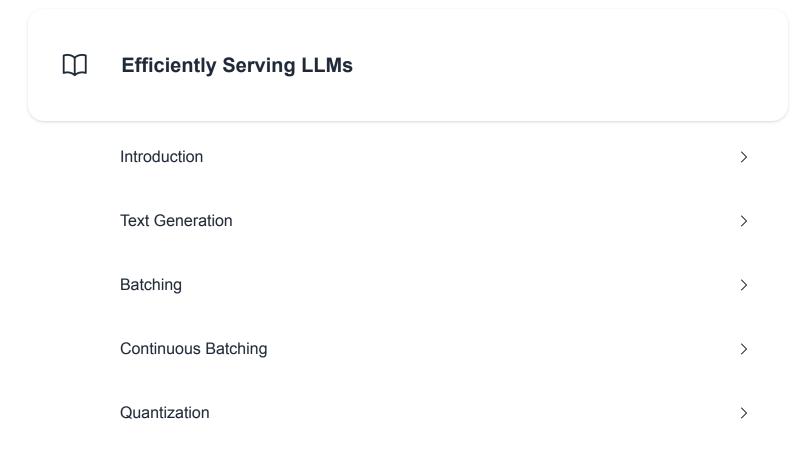






Lou Spironello, congratulations on completing Efficiently Serving LLMs!



Low-Rank Adaptation	>
Multi-LoRA inference	>
LoRAX	>
Conclusion	>