

Introduction

1. Tell me about yourself.
2. Explain your current GenAI/RAG project.
3. Which RAG architecture did you use?

RAG & Evaluation

4. How do you validate which RAG approach works better?
5. How do you evaluate the complete RAG pipeline?
6. Which evaluation metrics do you use?
7. How do you measure hallucinations?

Data & Security

8. Was your data public or confidential?
9. Is the RAG pipeline different for public and confidential data?
10. Suppose an HR salary spreadsheet is uploaded. Will you allow it? What strategy will you follow?
11. How do you implement access control in RAG?
12. How do you prevent unauthorized users from accessing confidential documents?

Classical Machine Learning

13. Have you worked on classical Machine Learning?
14. Which ML algorithms have you worked on?
15. Which algorithm did you use in your project?
16. Why did you choose that algorithm?
17. How did you evaluate your ML model?

Coding

18. Find the second largest element in an array.
19. Do it without sorting.
20. Do not use built-in functions.
21. Explain your approach.
22. How will you handle duplicate values?
23. What are the time and space complexities?

Deployment & Cloud

24. Did you deploy the application?
25. Explain your deployment architecture.
26. Which Azure services did you use?
27. How did you deploy the Docker container?
28. Where did you store secrets and API keys?

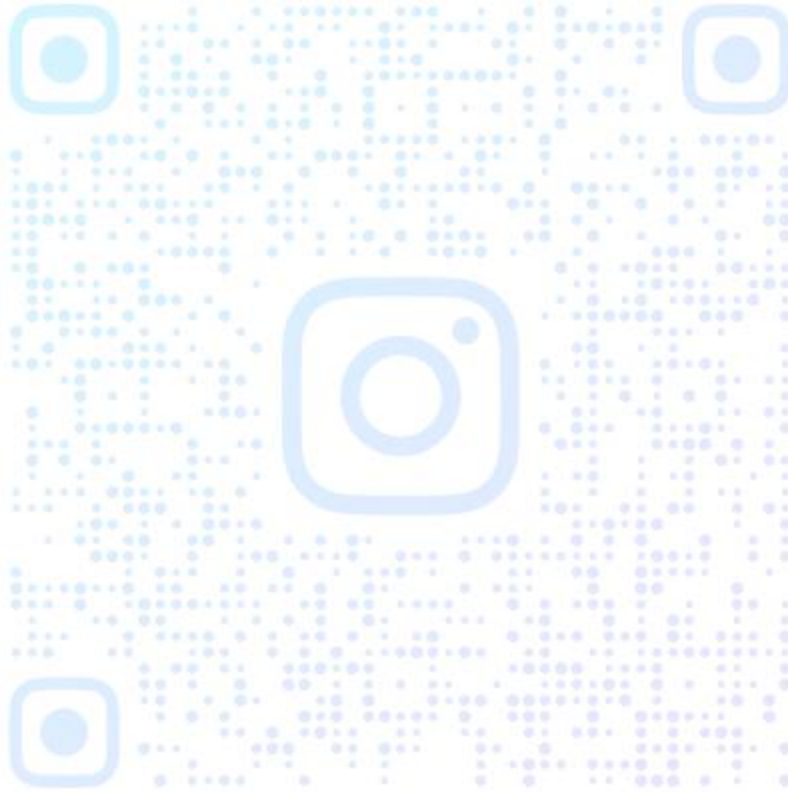
Performance Optimization

29. If you want to improve the performance of your application, where will you check?
30. How do you optimize RAG latency?
31. How do you optimize LLM response time?

<https://www.instagram.com/rawhul213>

32. Where would you use caching?

33. How do you identify the bottleneck?



@RAWHUL213