Feeling stuck in the AI / GenAI learning loop? 🔁 🤯

- Vibe coding is fine. Building AI apps with LLMs without deep knowledge is fine too, most of us do it to speed up development.
- But can you answer these fundamental AI/GenAI questions without help (if focusing on broader AI roles)?
- If you can’t answer most, strengthen your fundamentals along with hands-on projects. From my experience, many candidates struggle with core concepts in interviews.
- Most AI updates are currently just incremental (except for a few breakthroughs). Without strong fundamentals, you'll just be chasing AI in frustration.

Other Useful GenAI Resources

- Essential Terms GenAI
- GenAI Tech Stack
- GenAI Interview Q&A
- Agentic-AI Interview Q&A

Q1. Are you atleast familiar with traditional ML algorithms used for classification, regression, clustering and ensemble methods (e.g. decision trees, support vector machines, logistic regression, k-means, random forests, XGBoost etc)?

Q2. Do you have a understanding of probability, statistics, linear algebra and calculus enough to read and interpret research papers in AI?

Q3. Are you comfortable with the fundamentals of deep learning, including neural network architectures, backpropagation, optimization techniques and loss functions?

Q4. Can you explain how RNN-based models work and discuss their limitations (e.g. vanishing gradients, difficulties with long-term dependencies) compared to transformer architecture?

Q5. Can you describe the internal workings of transformer architectures, including the role of attention mechanisms and how they overcome RNN limitations?

Q6. Are you aware of the differences between encoder-only models (e.g. BERT), decoder-only autoregressive models (e.g. GPT, Claude, DeepSeek type), and encoder–decoder models (e.g. T5, BART), and can you provide examples of when each is used?

Q7. Why do most modern LLMs adopt decoder-only autoregressive architectures and what are the trade-offs compared to other architectures?

Q8. Can you explain the functions of the embedding layer, self-attention, cross-attention and multi-head attention in transformer architectures?

Q9. What role do linear (feedforward) layers play within transformer blocks and how do activation functions (e.g. ReLU, GELU) and softmax (in attention scoring) contribute to the model's performance?

Q10. What are tokens and embeddings and how do they serve as the primary inputs for language models?

Q11. Can you explain what raw logit scores are and how decoding parameters such as temperature, top-k and top-p (nucleus sampling) affect response generation in LLMs?

Q12. What is model quantization and what are the different quantization techniques (e.g. post-training quantization, quantization-aware training) used to optimize models for deployment?

Q13. Beyond quantization, how do techniques like pruning, knowledge distillation and low-rank approximations help in reducing model size and improving efficiency for deployment?

Q14. What is RAG, what techniques are available (including how to evaluate them), and how do they enhance model performance?

Q15. How do vector databases store and retrieve data and what is the significance of data chunking and embedding source data for improving retrieval quality?

Q16. What strategies exist for prompt optimization and evaluation in generative AI and how do different techniques (e.g. prompt chaining, prompt templates) influence model outputs?

Q17. How do you determine when to apply simple prompt chaining, RAG-based methods, or agent-based workflows (including autonomous agents) in various application scenarios?

Q18. What are the fundamentals of fine-tuning large models and what are the various supervised fine-tuning approaches (e.g. full model fine-tuning, adapter tuning, LoRA, QLoRA etc)?

Q19. Do you understand the basics of reinforcement learning and its application in language model training, including methods such as PPO, DPO, GRPO and Reinforcement Learning from Human Feedback (RLHF)?

Q20. What are the common evaluation metrics for machine learning, deep learning and generative AI models?

Q21. What are the key optimization algorithms (e.g. gradient descent, stochastic gradient descent, ADAM) and how do various loss functions (such as cross-entropy or mean squared error) impact model training?

Q22. How does transfer learning work in deep learning and what strategies are effective for adapting pre-trained models to new domains or tasks?

Q23. Can you explain different methods of knowledge distillation, including soft distillation (using logit probability distribution) and hard distillation (using hard labels)?

Q24. Do you know at least basic of MLOPs/LLMOPs process ?

AI has many other fundamentals. In my experience, if you're comfortable with these, keeping up with AI’s rapid updates becomes much easier. Like it or not, these fundamentals will surface many times in your interviews.