Top Interview Questions & Best Possible Answers

1. **What problem does your OCR project solve?**

The project solves the issue of non-searchable scanned PDFs, which are basically images. Many users cannot search names, text, or keywords inside such PDFs. I personally faced this when someone asked me to find a name inside an image-based PDF and search didn’t work. So I built a tool that converts normal PDFs into fully searchable, text-embedded PDFs using OCR—totally free, lightweight, and offline.

2. **Why did you choose Tesseract OCR?**

I chose Tesseract OCR because it is open-source, highly accurate for document OCR, supports multiple languages, and integrates easily with Python using pytesseract. It also provides a direct function image_to_pdf_or_hocr() which lets me embed OCR text directly into PDF, keeping layout intact.

3. **How does your tool work internally?**

- Extracts images from each page of the PDF
- Uses pytesseract + custom OCR config to detect text
- Generates text-layer-embedded PDF pages
- Merges all pages into a single searchable PDF
- Outputs the final file offline

4. **What challenges did you encounter?**

- Tesseract PATH issues
- DPI tuning for accuracy
- Layout preservation
- Packaging into .exe
- Missing Poppler/GhostScript

5. **Why create a .exe?**

To make it accessible for non-technical users without needing Python or installations.

6. **Libraries used:**

pytesseract, pdf2image, Pillow, PyMuPDF, io, PyInstaller.

7. **Future improvements:**

Multi-language support, batch processing, format options, UI themes, Mac/Linux builds, MSI installer.

8. **How do you ensure accuracy?**

High DPI (300), custom config, preprocessing, correct PSM modes.

9. **Most interesting learning?**

Understanding OCR engines, PDF layers, bundling .exe, dependency handling.

10. **How would you scale?**

GUI app, Web API, Docker, GPU OCR.

11. **How is your tool different?**

Free, offline, privacy-safe, lightweight, high-quality output.

12. **If you had more time?**

Tkinter GUI, drag-drop, batch mode, cloud OCR, logging, MSI installer.