■ OCR Research Paper – Interview Cheat Sheet

Elevator Pitch:
“This research paper compares traditional OCR (like Tesseract) with transformer-based models (Donut, LayoutLMv3, TrOCR) for structured and semi-structured document understanding in banking, healthcare, and insurance. We benchmarked models on the FUNSD dataset using metrics like CER, WER, F1-score, BLEU, and string similarity. The study shows that while traditional OCR like Tesseract still performs decently on structured forms, transformer-based models like Donut and LayoutLMv3 excel in semantic and contextual understanding, though at higher computational costs. Our results suggest hybrid models as a future direction.”

Workflow:
1. Dataset & Preprocessing (FUNSD dataset, denoising, binarization, CLAHE, deskewing).
2. Models Compared (Tesseract, Donut, LayoutLMv3, TrOCR).
3. Evaluation Metrics (CER, WER, F1-score, BLEU, String Similarity).
4. Results (Donut best semantically, TrOCR best for handwriting, Tesseract good for structured forms).

Model	Working	Strengths	Weaknesses	Results
Tesseract	Segmentation + OCR pipeline (LSTM-based)	Lightweight, good for printed structured forms	Fails on handwriting, semi-structured layouts	CER = 0.65, F1 = 0.55 (improved slightly with preproces
Donut	Vision-to-Text Transformer (Swin Encoder + Decoder)	End-to-end, semantic accuracy, multilingual	Requires GPU, higher compute	BLEU = 0.82, String Similarity = 0.89 (best semantic
LayoutLMv3	Multimodal: text + layout + visual embeddings	Strong for structured forms, captures spatial info	Overfits, poor on headers	+83% on answers, -51.5% on headers
TrOCR	Vision-Language Pretraining (OCR Transformer)	Excellent for handwritten text, lower CER	Heavy compute requirements	≈40% lower CER vs Tesseract on handwriting

Key Findings:
- Tesseract: Best for structured forms but limited in flexibility.
- Donut: Best semantic understanding.
- LayoutLMv3: Strong on entity extraction but overfits.
- TrOCR: Best for handwritten forms.

Conclusion: No single best model → Hybrid pipelines + preprocessing are the future.