

## GURASEES SINGH RAI

Phone: (408)-680-7324 | Email: [guraseesrai2@gmail.com](mailto:guraseesrai2@gmail.com) | [LinkedIn](#) | [GitHub](#)  
Newark, CA

### SUMMARY

- Data Engineer & Applied AI Practitioner with 1+ year of research and industry internship experience, specializing in designing end-to-end data pipelines, deploying ML models into production, and scaling solutions for large structured/unstructured datasets.
- Hands-on experience across RNA sequence modelling with LSTMs and FastAPI, NLP-driven document processing using LangChain + GPT-4, housing price prediction (scikit-learn), e-commerce customer analytics, and video deepfake detection.
- Proficient in Python, SQL, PyTorch, scikit-learn, and cloud platforms (Azure, AWS, GCP), with a proven ability to translate complex technical workflows into actionable business insights.

### WORK EXPERIENCE

#### *Oregon State University, OR | AI Research Intern*

*Aug 2025 - Present*

- Built preprocessing pipeline for viroid RNA cut-site prediction: FASTA/CSV ingestion, leak-free splits, and one-hot encoding of nucleotide windows.
- Designed baseline ML models and trained LSTMs, tracking AUROC and PR-AUC for rigorous evaluation.
- Exposed trained models via FastAPI endpoints to score new RNA sequences in real time.

#### *Oregon State University, OR | Applied AI Research Assistant*

*Feb 2025 - Jun 2025*

- Developed LangChain + GPT-4 document processing pipeline transforming PDF/Word learning materials into structured educational scripts using PyPDFLoader, RecursiveCharacterTextSplitter (1000-character chunks, 100-character overlap), and ChatOpenAI with temperature=0.7.
- Implemented intelligent token management system using tiktoken encoding with 1500-token limits for context and script generation, including dynamic text truncation functions (truncate\_to\_token\_limit()) to optimize GPT-4 API calls and prevent token overflow.

#### *Techlive Solutions, India | Data Analyst Intern*

*Jan 2022 - Jun 2022*

- Engineered scikit-learn preprocessing pipeline (imputation, scaling, encoding) to support a housing price prediction model.
- Created custom ratio-based features improving interpretability (e.g., rooms\_per\_household, bedrooms\_per\_room).
- Achieved RMSE of ~\$66.9K on test set using stratified train/test split and validated findings with visual EDA and correlations.

### PROJECTS

#### *E-Commerce Customer Analytics | [GitHub Link](#)*

- Developed end-to-end ML pipeline processing 541K transactions across 38 countries; implemented RFM segmentation achieving silhouette score 0.617 and Random Forest classifier with 99.42% accuracy.
- Built production-ready customer intelligence system: comprehensive data preprocessing, feature engineering, K-means clustering validation, and cross-validated classification model.

#### *Hybrid Deepfake Detector | [GitHub Link](#)*

- Developed ResNet3D + Swin Transformer pipeline for video-based deepfake detection with temporal modeling through transformer encoder layers.
- Built comprehensive video processing system: frame extraction, data caching, model training with BCEWithLogitsLoss, and performance evaluation with multiple metrics.

### SKILLS

- **Programming/ML:** Python, SQL, PySpark, Pandas, NumPy, Data Pipelines.
- **Machine Learning & AI:** PyTorch, scikit-learn, TensorFlow, LangChain.
- **Cloud Technologies:** AWS (S3), Azure ML, GCP Vertex AI.
- **Visualization:** Matplotlib, Seaborn, Plotly, Streamlit.
- **Version Control & Databases:** GitHub, Git, SQL Server, PostgreSQL, MongoDB, MySQL.
- **Operating Systems:** Windows, macOS, Linux (Debian).

### EDUCATION

#### **Master of Engineering in Computer Science (GPA – 3.98)**

Oregon State University, Corvallis, OR

*Sep 2023 - Jun 2025*

#### **Bachelor of Engineering**

Thapar Institute of Engineering & Technology, India

*Aug 2018 - Jun 2022*

### CERTIFICATIONS

- **Data Analytics and Machine Learning Certificate**  
ElevateMe Bootcamp
- **IBM Data Engineering Certificate**  
Coursera

*May 2025 - Aug 2025*

*Mar 2025 - May 2025*