software engineering
@ UMD

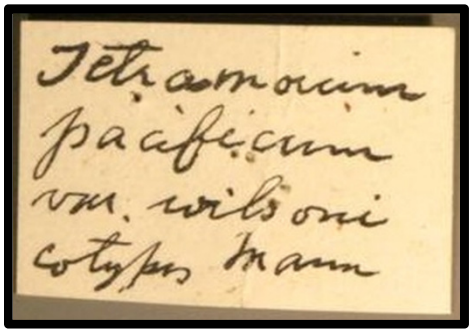
Specimen Transcription

Aidan Caruso, Samuel Cleary, Anuraag Karunakaran
Trevor Moore, Maiwand Raheem

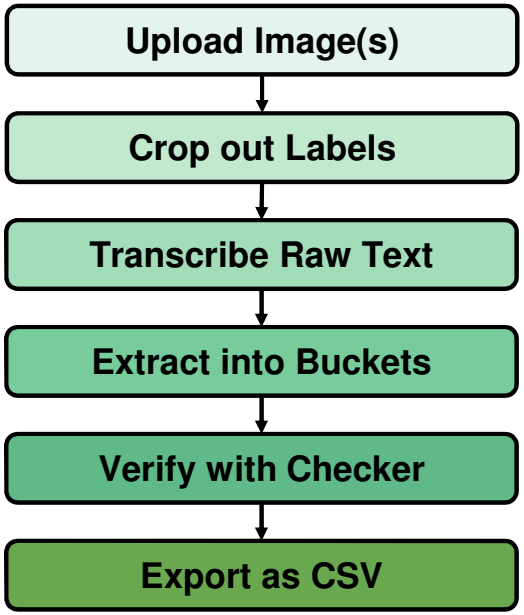
COLLEGE OF COMPUTER, MATHEMATICAL, & NATURAL SCIENCES

Problem Definition

- The Smithsonian spends **$100,000+** & days of work time on manual transcriptions each year.
- **Automate** and **accelerate** this process using AI and LLMs.



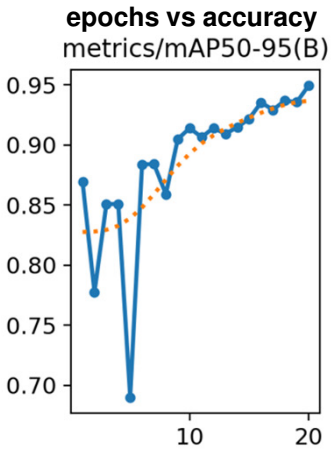
Final Design



Design Analysis

1.YOLO v8

- Reaches **95% accuracy**, errors are duplicates which don't affect the LLM
- Tests on non insect labels showed **90%+ accuracy** with drops caused by an increase in duplicates.



2. LLMs

Model	OCR	gemma3 text	gemma3 buckets
Accuracy	<30%	82.2%	77%

- **Traditional Optical Character Recognition(OCR)** models are bad with hand written, blurred, and smudged text.
- **Large Language Models** use context to guesstimate the closest word based off linguistics and a scientific knowledge base.

Prototype Testing

8 Tests Conducted

Testing Procedures

Each tester does 3 rounds of testing with the product:

1. **Unguided test** run with provided image set-1
2. **Guided test** run with provided image set-2
3. **Open testing** with images chosen by testers

References

- Smithsonian Institution. (n.d.). *Entomology Department.* https://naturalhistory.si.edu/research/entomology
- iDigBio. (2015, April 6). *Simultaneous transcription blitzes a success!* iDigBio. https://www.idigbio.org/content/simultaneous-transcription-blitzes-success