

# **Automatic Test Generation With Codex Using Natural Language Artifacts**

---

Aakash Reddy Bhoomidi

Tejas Kumar Leelavathi

Varsha Chandrahasareddy Mulangi

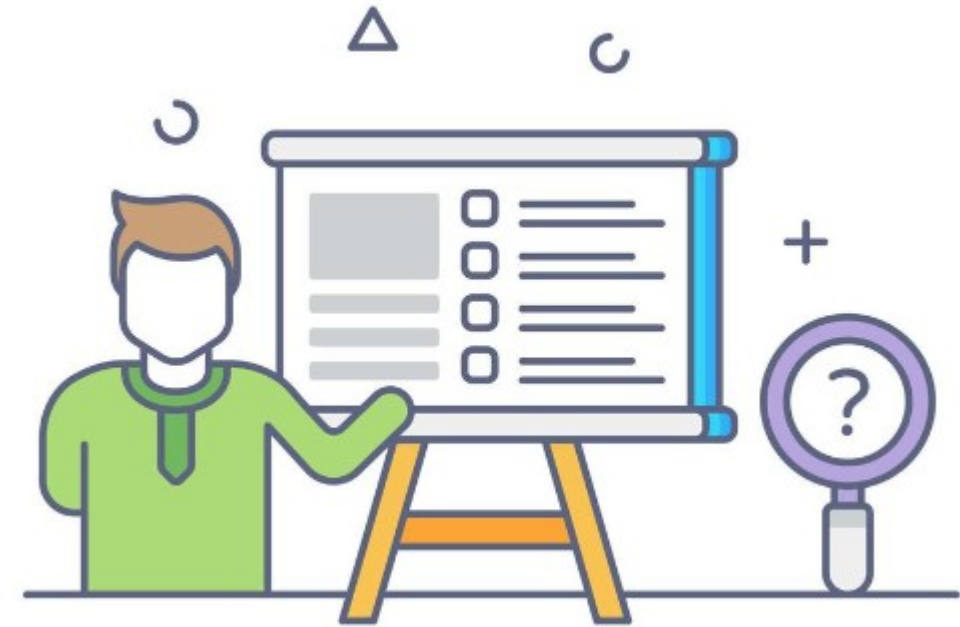
Veda Varshitha Oruganti



# Problem Statement

---

- Our research project focuses on evaluating and enhancing the performance of Open AI's Codex model in generating test cases based on requirements expressed in natural language.
- The objective is to cover a wide range of scenarios, including edge cases, with minimal human intervention.



# Why?

---

- **Efficiency and Speed**
- **Resource Optimization**
- **Cost Savings and Benefits**
- **Quality Assurance**



# Research Questions

---

- How effectively does Codex translate natural language requirements into functional test cases?
- Does fine-tuning Codex on a specific dataset improve its ability to generate relevant and accurate test cases?
- Can the fine-tuned Codex model effectively generalize its capabilities to new requirement documents not included in the training data?



# Design

---

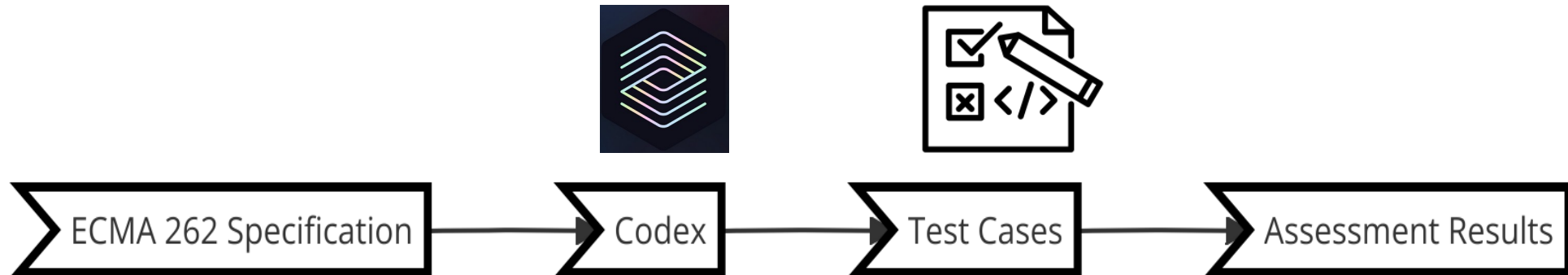
- We have divided our proposed design into 3 different sections answering our 3 Research Questions:



# Section 1

---

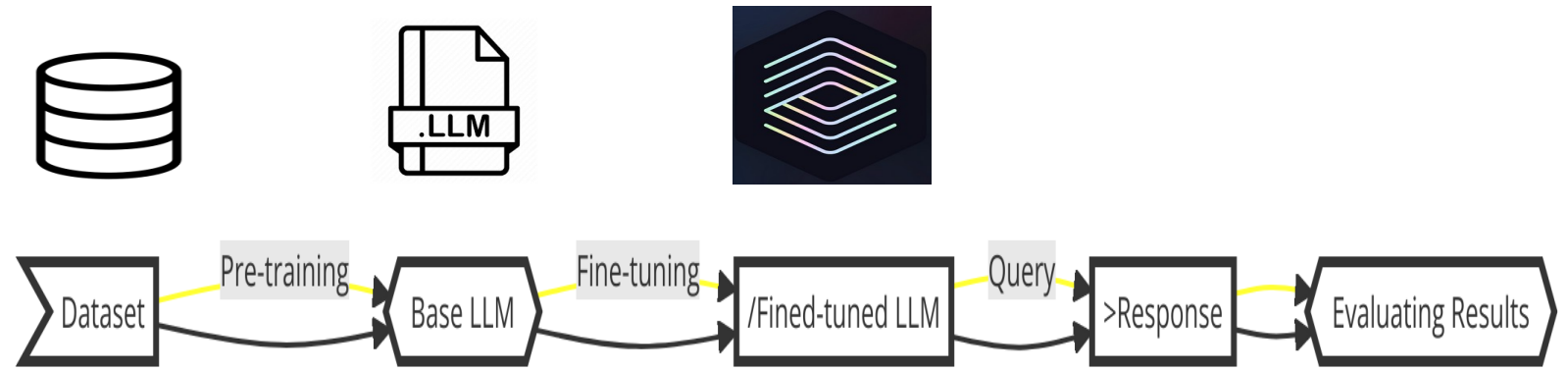
- Collect natural language requirements from the ECMA 262 specification.
- Utilize Codex to create test cases based on these requirements.
- Assess the generated test cases for their completeness, accuracy, and how well they adhere to the original requirements from ECMA 262.



# Section 2

---

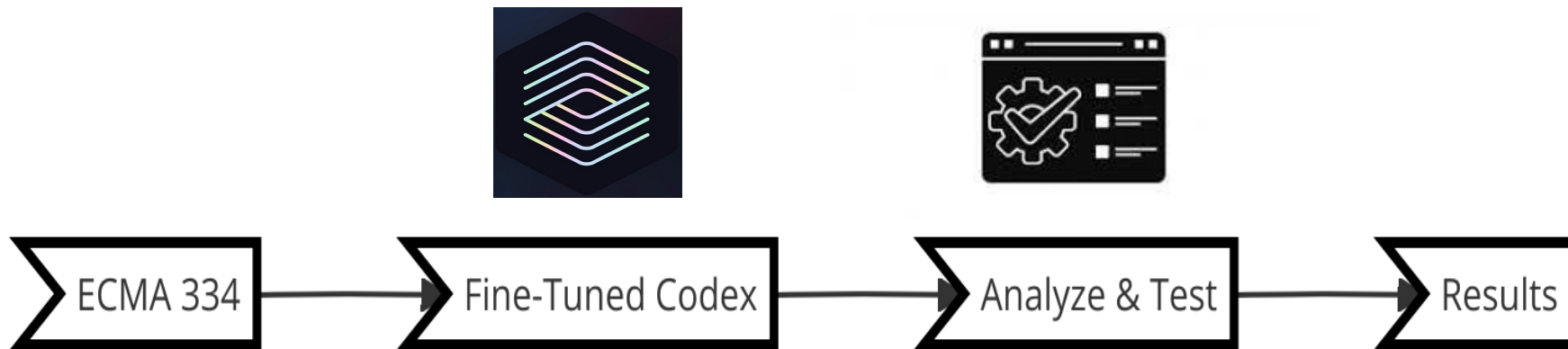
- Assemble a dataset comprising natural language descriptions of ECMAScript (JavaScript) requirements and corresponding test cases focusing on boundary and exceptional cases
- Fine-tune Codex on this dataset
- Compare the performance of the fine-tuned model against the baseline (pre-fine-tuning) model



# Section 3

---

- Test the fine-tuned Codex model on a new set of requirements that were not part of the training or fine-tuning process.

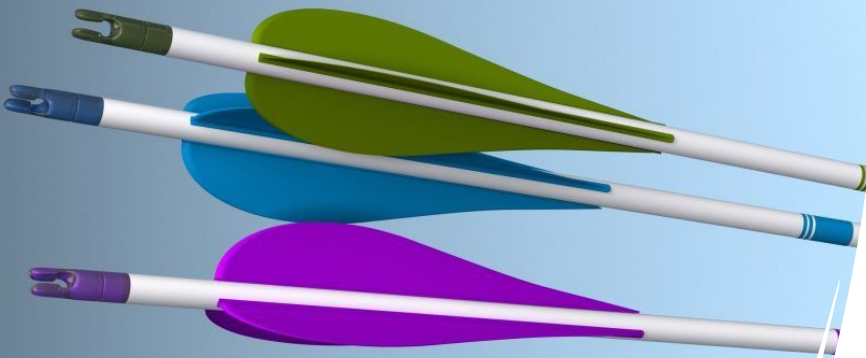




# Evaluation

---

- Compare the performance of the fine-tuned model against the baseline (pre-fine-tuning) model.
- Assess the quality of the generated test cases in terms of relevance, accuracy, and completeness.
- Generalization capability, accuracy on new data, and any reduction in the need for manual correction.



# Planned Timeline

