King Saud University
College of Computer and Information Sciences

Department of Information Technology

جامعة
الملك سعود
King Saud University

كلية علوم الحاسب والمعلومات
قسم تقنية المعلومات

---

# IT362: Principles of Data Science

# Phase 1: Data Collection, Research, and Assessment

Prepared By

| Section | Student Name | ID |
|---|---|---|
| 80657 | Afnan Alkharji | 443200897 |
| | Deema Alfarhoud | 444200955 |
| | Hussah Alotaibi | 444201039 |
| | Falwa alkhalifah | 444200745 |
| | Sana Alotaibi | 444200426 |

Supervised by:

Dr. Abeer Aldayel

## Contents

## Introduction

The rise of large language models, such as ChatGPT, has significantly impacted the field of artificial intelligence. This project investigates whether this shift has also affected the programming languages used in open-source AI development.

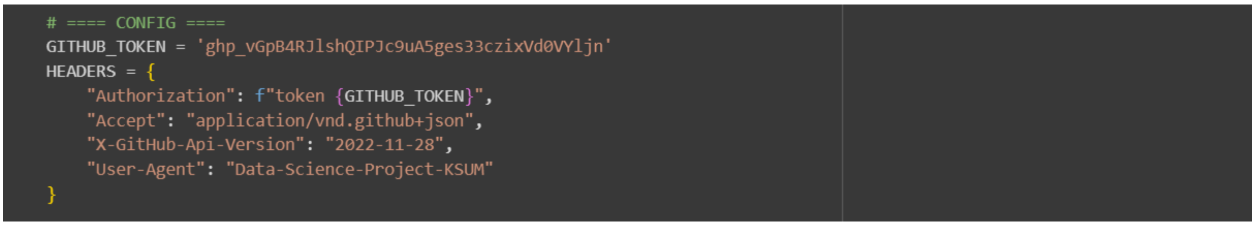
By collecting and analyzing metadata from popular AI-related GitHub repositories before and after the release of ChatGPT, we aim to examine trends in language usage over time. The goal is to determine if the growth of generative AI has influenced developers' preferences for programming languages within the AI community.

## Data sources

### Data Source

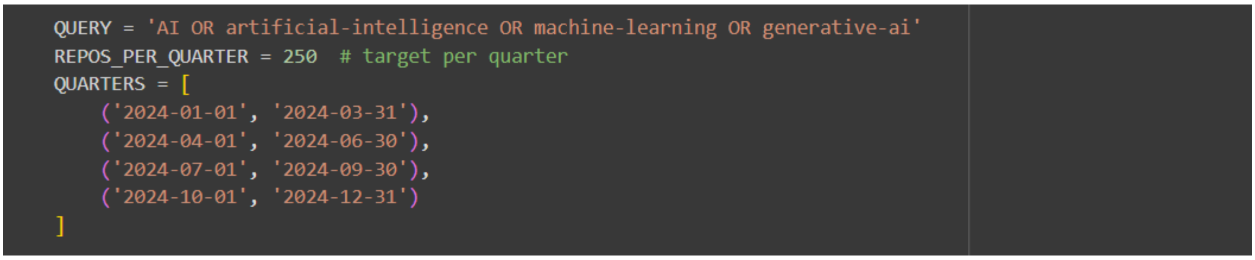
The data source for this study is the GitHub REST API. Which provides a programmatic interface to access repositories information.

### Data Collection



```
# ==== CONFIG ====
GITHUB_TOKEN = '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'
HEADERS = {
    "Authorization": f"token {GITHUB_TOKEN}",
    "Accept": "application/vnd.github+json",
    "X-GitHub-Api-Version": "2022-11-28",
    "User-Agent": "Data-Science-Project-KSUM"
}
```

The dataset will be collected over a period spanning six years, from 2020 to 2025. For each year, we will identify and collect the top 1000 most-starred AI related repositories distributed over quarters (250 per quarter) to guarantee we capture the trends over the whole year.



```
QUERY = 'AI OR artificial-intelligence OR machine-learning OR generative-ai'
REPOS_PER_QUARTER = 250  # target per quarter
QUARTERS = [
    ('2024-01-01', '2024-03-31'),
    ('2024-04-01', '2024-06-30'),
    ('2024-07-01', '2024-09-30'),
    ('2024-10-01', '2024-12-31')
]
```

This means we collected approximately 1000 observations for each year, except for 2025 where we collected repositories up to 23rd of September and it resulted in 754 repositories. This will result in a total of approximately 5751 observations for our analysis.

### Data Features

The following features will be collected for each repository:

| Feature | Data type | Description |
|---|---|---|
| name | Qualitative (Nominal) | Name of the repository |
| owner | Qualitative (Nominal) | User or organization that owns the repository. |
| url | Qualitative (Nominal) | URL of the repository |
| description | Qualitative (Nominal) | A brief description of the project |
| stars | Quantitative (Ratio) | Number of times the repository has been starred |
| forks | Quantitative (Ratio) | Number of times the repository has been forked |
| created_at | Qualitative (Ordinal) | Date of the repository's creation |
| pushed_at | Qualitative (Ordinal) | Date of the last update or push to the repository |
| primary_language | Qualitative (Nominal) | Main programming language of the repository |
| all_languages_bytes | Qualitative (Nominal) | A dictionary showing the breakdown of all programming languages used in the repository and the number of bytes for each |
| topics | Qualitative (Nominal) | A list of topics or tags associated with the repository |
| contributors_count: | Quantitative (Ratio) | Number of unique contributors to the repository |

**Dataset Bias Evaluation**

- **Representation Bias:** By focusing on the "most-starred" repositories, our dataset represents projects that have gained most attention, yet their fame could be due to effective marketing, or a high-profile creator and not necessarily trendy.
- **Measurement Bias:** By dividing trends by quarter, we ensure each period is represented. However, a "smaller" trend in an active quarter could be more impactful than a "main" trend in a less active quarter.
- **Historical Biases:** By analyzing existing repositories, we could be missing projects that were more popular during their time but have since been removed.

## Objectives

Our project aims to uncover "**how the emergence of large language models (LLMs), such as ChatGPT, has influenced programming language preferences in open-source AI development**." Specifically, the objectives are to:

1. **Analyze Temporal Trends**
   Examine changes in the frequency and proportion of programming languages used in popular AI-related GitHub repositories before and after the release of ChatGPT.
2. **Identify Shifts in Developer Preferences**
   Detect any significant increases or decreases in the use of specific languages (e.g., Python, Java, Julia, R, C++) that may be associated with the rise of generative AI.
3. **Compare Pre- and Post-ChatGPT Periods**
   Statistically compare programming-language usage in the two time windows to evaluate whether the advent of LLMs correlates with measurable changes in developer behavior.
4. **Characterize the AI Development Ecosystem**
   Provide a descriptive overview of the most commonly used languages, their growth or decline rates, and how these patterns reflect evolving practices in AI research and engineering.
5. **Generate Insights for Future AI Tooling**
   Offer evidence-based insights that can guide educators, practitioners, and technology strategists in understanding which programming languages are gaining traction in the era of generative AI.

## Method

The dataset collected from GitHub will be used to analyze trends in programming languages among AI-related projects over multiple years (2020–2025). Each repository includes information such as primary programming language, all languages used, number of stars, forks, topics, and contributors.

Using this dataset, the following steps will be taken:

1. **Data Cleaning & Preprocessing**: Ensure all entries have consistent formats, handle missing values, and normalize language names if needed.
2. **Descriptive Analysis**: Calculate KPIs such as the most common programming languages per year, average stars per language, and number of AI-related repositories created annually.
3. **Trend Analysis**: Compare language popularity across years to identify shifts in which languages are used in AI projects over time.
4. **Visualization**: Create charts and graphs to illustrate language trends, repository activity, and other key metrics.
5. **Insights**: Draw conclusions about how AI development trends have influenced programming language adoption year by year.

## Challenges faced in data collection and recommendations

**Challenges:**

1. **API Rate Limits**: GitHub restricts the number of requests per hour. Without a valid token, the script could fail if too many requests are made.
2. **Time-Consuming Collection**: Fetching detailed data for many repositories takes several minutes per quarter.

**Recommendations:**

- Always use a valid personal access token to avoid hitting rate limits.
- For future projects, consider storing raw data snapshots at regular intervals to avoid repeating long collection processes.