40th IEEE/ACM International Conference
on Automated Software Engineering (ASE 2025)

PyGress: Tool for Analyzing the Progression of Code Proficiency in Python OSS Projects

Rujiphart Charatvaraphan, Bunradar Chatchaiyadech, Thitirat Sukijprasert, Chaiyong Ragkhitwetsagul, Morakot Choetkiertikul, Raula Gaikovina Kula, Thanwadee Sunetnanta, Kenichi Matsumoto

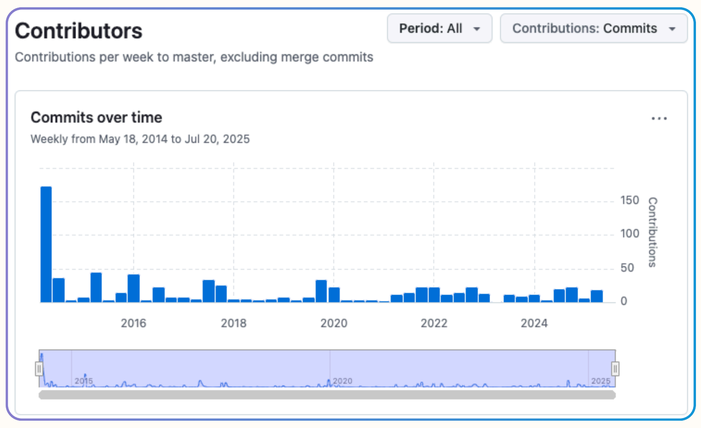
Faculty of Information and Communication Technology, Mahidol University, Thailand
Graduate School of Information Science and Technology, The University of Osaka, Japan
Graduate School of Information Science, Nara Institute of Science and Technology, Japan

ABSTRACT

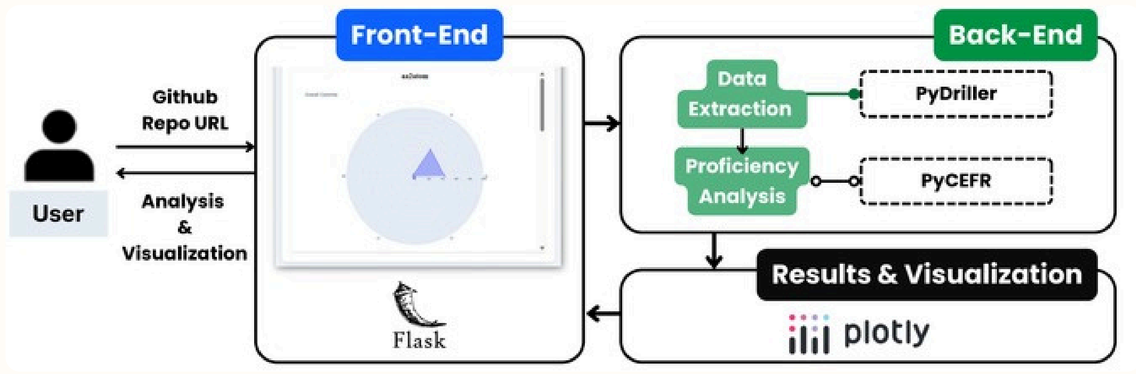
Assessing developer proficiency in open-source software (OSS) projects is essential for understanding project dynamics, especially for expertise. This paper presents **PyGress**, a web-based tool designed to automatically evaluate and visualize Python code proficiency using pycefr, a Python code proficiency analyzer. By submitting a GitHub repository link, the system extracts commit histories, analyzes source code proficiency across CEFR-aligned levels (A1–C2), and generates visual summaries of individual and project-wide proficiency. The PyGress tool visualizes per-contributor proficiency distribution and tracks project code proficiency progression over time. PyGress offers an interactive way to explore contributor coding levels in Python OSS repositories.

PROBLEM STATEMENT

- No existing tool evaluates and visualizes Python code proficiency in OSS projects.
- Current studies focus on contribution volume or activity, not coding competency.



SYSTEM ARCHITECTURE



FRAMEWORK (PYCEFR)

What is CEFR?

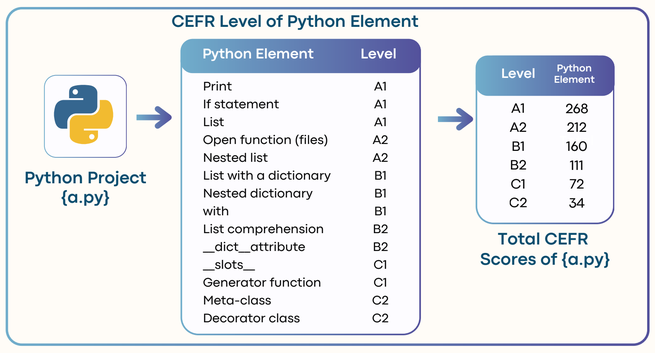
- An international standard for assessing language proficiency across six levels: A1–C2

Level Group	Level
A Basic	A1: Breakthrough or beginner
	A2: Waystage or elementary
B Independent	B1: Threshold or intermediate
	B2: Vantage or upper intermediate
C Proficient	C1: Effective operational proficiency
	C2: Mastery or proficiency

CEFR (Common European Framework of Reference for Languages)

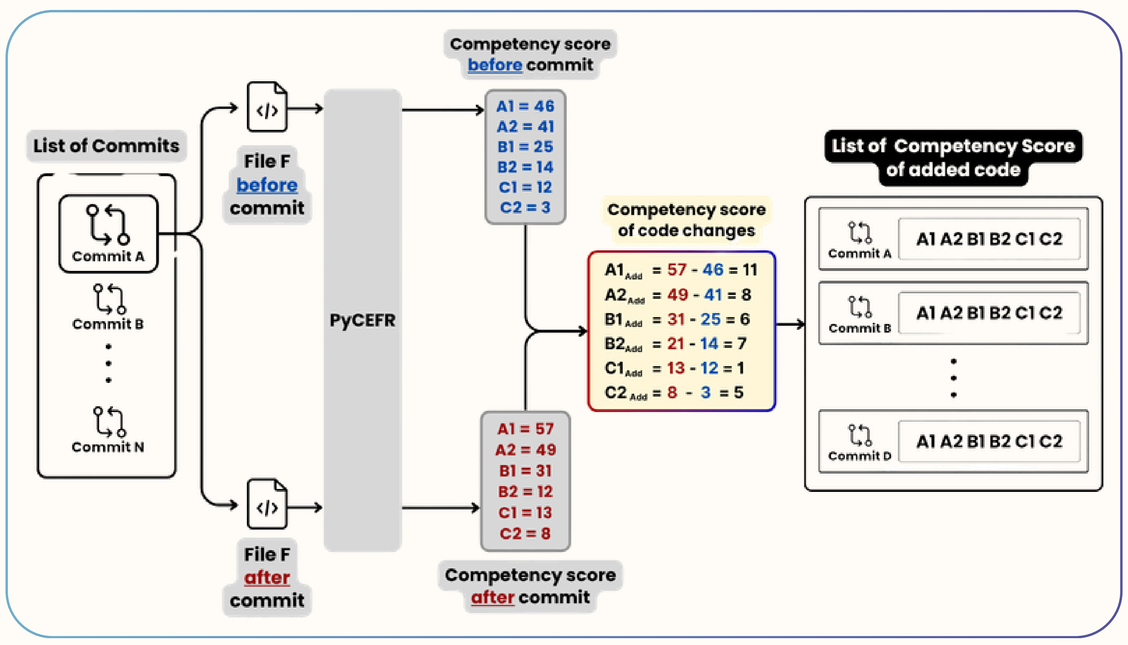
What is PyCEFR?

- A tool that categorizes Python code into 6 proficiency levels, aligning with the **CEFR (A1, A2, B1, B2, C1, and C2)**.



Python Project's CEFR Level Calculation (PyCEFR)

PROFICIENCY ANALYSIS



1) Repository Submission

- Users provide a GitHub repository link to PyGress. The tool automatically retrieves the **commit history** and associated **source code files**.

2) Code Proficiency Analysis

1. Python Code Extraction and Preprocessing
2. Proficiency Assessment Using PyCEFR
3. Calculating Proficiency Scores Per Contributor Over Time
4. Aggregating Project-wide Scores

3) Visualization and Interaction

- **PyGress generates interactive graphs for easy exploration:** Per-contributor proficiency distribution and Project-wide proficiency progression over time

FEATURES OF PYGRESS

Python Code Proficiency Evaluation

- Automatically analyzes Python code from GitHub repositories.
- Uses PyCEFR to classify code constructs into CEFR levels. (A1 to C2), reflecting developer skill levels.

Contributor and Project-Level Tracking

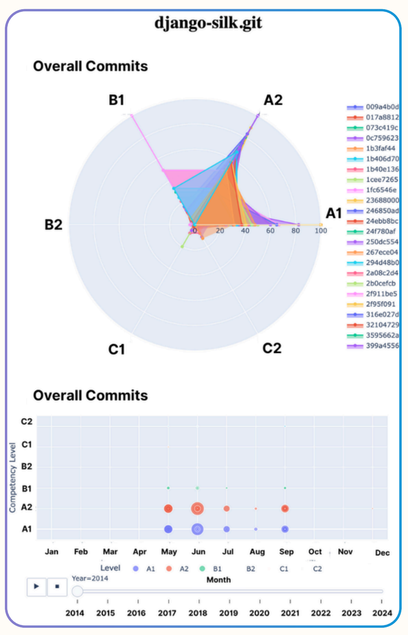
- Tracks proficiency progression for each contributor over time.
- Aggregates data to show project-level proficiency evolution.
- Helps identify key contributors and potential risks.

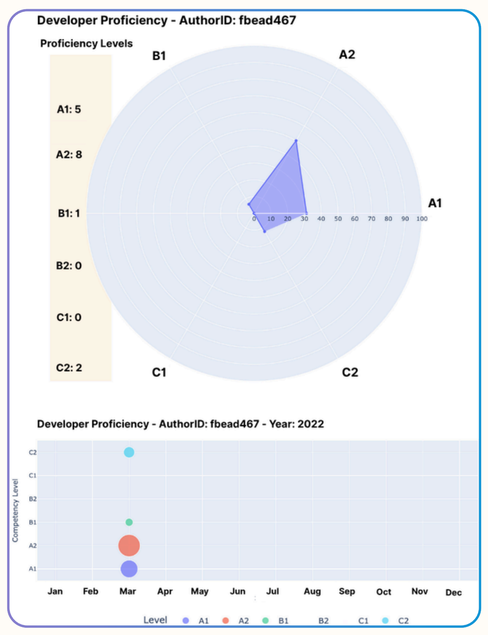
Interactive Visualization Interface

- Provides dynamic charts (spider and slider graphs) via a web-based UI.
- Enables maintainers to explore proficiency data easily and make informed decisions.

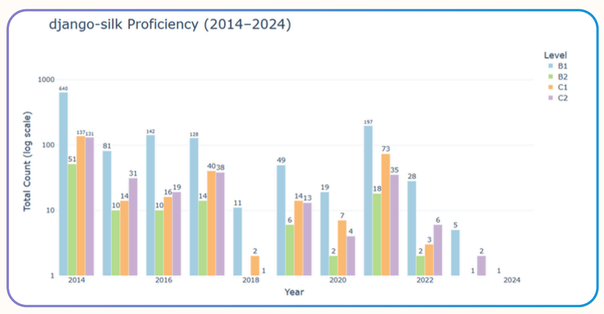
RESULTS

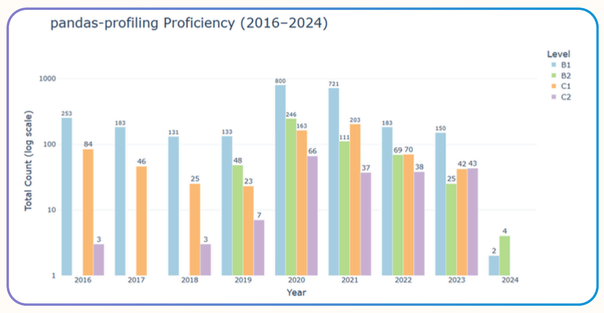
DJANGO-SILK: INDIVIDUAL CONTRIBUTOR'S PROFICIENCY





PYTHON PROFICIENCY PROGRESSION OF 2 OSS PROJECTS





DEMO & OTHERS

Pre-print:
Scan to get the pre-print of the paper on arXiv

Video demo:
Scan to watch the PyGress demo video

GitHub repo:
Scan to view the PyGress source code on GitHub