NLP701 and NLP805: Fall 2025
Assignment 2: Detecting AI-Generated Code

Preslav Nakov, Ted Briscoe
MBZUAI
{preslav.nakov,ted.briscoe}@mbzuai.ac.ae

Abstract

- This is an *individual* assignment.
- This assignment carries 10 points.
- Deadline: November 28, 2025 (23:59 UAE time).
- Deliverables: (*i*) your code, (*ii*) a file with your predictions on public test set (you can just share a copy of sample submission with your predictions), (*iii*) your ID from the SemEval-2026 task 13 leaderboard, and (*iv*) a report (1-2 pages).
- All required material should be zipped in one folder and submitted on Moodle.

1 Detection of AI-generated Code

The assignment involves developing systems to identify code, generated by LLMs. This is essential for preventing misconduct in diverse areas (programming competitions, assignments, etc.). Such systems can be integrated into Turnitin, VS code, or GitHub extensions (or can be used to test your submissions this very assignment).

2 Your Task

You may choose one or more of the following tasks:

- **Subtask A** (Binary Machine-Generated Code Detection Task): Given a code snippet, predict whether it is *fully human-written* or *fully machine generated*. The main difficulty of this task is that the training set contains code in a limited range of languages and only from one programming domain (Algorithmic problems), while the test set is diverse and has no such restrictions.

- **Subtask B** (Multi-Class Authorship Detection Task): Given a code snippet, predict its author which may be (i) Human, or (ii–xi) one of ten LLM families: DeepSeek-AI, Qwen, 01-ai, BigCode, Gemma, Phi, Meta-LLaMA, IBM-Granite, Mistral, OpenAI.

- **Subtask C** (Hybrid Code Detection Task): Classify each code snippet as (a) *Human-written*, (b) *Machine-generated*, (c) *Hybrid*, i.e., partially written or completed by LLM, or (d) *Adversarial*, i.e., generated via adversarial prompts or RLHF to mimic humans.

3 Important Notes

Here are some important things to consider:

- More details about each subtask can be found on the shared task website: https://github.com/mbzuai-nlp/SemEval-2026-Task13/

- Subtask A can be accessible via the following link: https://www.kaggle.com/t/99673e23fe8546cf9a07a40f36f2cc7e

- Subtask B can be accessible via the following link: https://www.kaggle.com/t/65af9e22be6d43d884cfd6e41cad3ee4

- Subtask C can be accessible via the following link: https://www.kaggle.com/t/005ab8234f27424aa096b7c00a073722

- You will have 50 submissions per day, but you can also do validation locally, since for the public set, the labels are available.

- You **can only use the TRAIN set to train your models**. Additional generations are prohibited by the competition rules. You cannot use the **TEST set or TEST SAMPLE** for training, since this is the public dataset. The labels in this dataset are given for validation only, and you **should not copy them into your submission**, since it is against the competition rules and can be considered as a misconduct.

- If you want to participate in this task after the assignment deadline, follow the mailing list

from the competition's GitHub, since the test set will be released in January 2026, and only predictions on this set will be considered for the final leaderboard.

- This time, you should **submit a zipped code base (repository)** (not a Jupyter notebook), as well as your predictions. The file format is described on the webpage of SemEval-2026 task 13 and on the competition's Kaggle page; it is the same format as for your required submission to the leaderboard.

- The evaluation measure for all subtasks is macro F-score.

4 Submission

For submission, include your full code along with a README that describes how to run your code, your predictions on the TEST set, and your team[1] name for the SemEval-2026 task 13 Leaderboard (feel free to use a name that can protect your identity from public discovery). The predictions format should be the same as for the SemEval-2026 task 13 leaderboard.

You also need to submit a report describing your approach, results, and analysis (1-2 pages) in standard *ACL format.[2] The report should include your Leaderboard team name, a detailed description of your features, evaluation results on the development set, and some analysis of how these features affected the performance of your models. Your submission should be zipped in one folder and submitted on Moodle.

5 Assessment

This is an individual assignment that carries ten points towards your final grade. Discussing the problems with each other is encouraged, but copying each others' code or report is strictly prohibited. You will be evaluated based on the richness of models and features explored, strength of results achieved, and quality of writing of your report.

[1] The SemEval-2026 task talks about teams, but for you this is an *individual* assignment.

[2] https://www.overleaf.com/read/crtcwgxzjskr