CS 2731
Introduction to Natural Language Processing

Session 16: Project proposal presentations

October 20, 2025

University of Pittsburgh | School of Computing and Information

Course logistics

- Next two class sessions will be guest lectures!
- Emma Jordan will be speaking this Wed Oct 22 on reinforcement learning
- Lorraine Li will be giving a guest lecture on her current NLP research on Mon Oct 27

Schedule

1. Raul, Shaojun, Naman
2. Charitha, Maria, Surabhi, Shubham, Victor, John
3. Uma, Yudan, Chase
4. Nate, Zhiwei, Hongbo
5. Tim, Keshav, Lucy

Instructions

- Plan for **7 min presentations max** not including Q&A
- Cover at least these key points
  - Project motivation (what is the value of this work?)
  - Briefly, what 1-2 other related papers have done (1 slide max)
  - What data you are planning to use
  - What approach/methods you plan to take
  - How you will evaluate your approach
- Put your slides in this presentation after your project name slide by class session, 2:30pm on Mon Oct 20

1. Raul, Shaojun, Naman

TextSlayer: A LLM-Based approach to play word adventure games

Naman Gupta, Raul Viteri, Shaojun Zheng

Introduction

- TextWorld: a Microsoft text-based game generation framework for NLP agent evaluation.
- Evaluates reasoning, memory, planning, and commonsense understanding.
- Simulates interactive decision-making scenarios unlike static benchmarks.
- Serves as a virtual test for real world scenarios for Embodied AI and Interactive AI

```
nag186@login1.crc.pitt.edu ssh ~/project/text_world
tw-play tw_games/custom_game.z8

Hey, thanks for coming over to the TextWorld today, there is something I need
you to do for me. First step, retrieve the American limited edition keycard from
the type 1 box. And then, unlock the American limited edition gate. After that,
open the American limited edition gate. Then, make an effort to move east.
Following that, pick-up the shirt from the floor of the attic. Got that? Good!

-= Scullery =-
You've stumbled into a normal room. Your mind races to think of what kind of
room would be normal. And then it hits you. Of course. You're in the scullery.
You begin to take stock of what's in the room.

You can make out a type 1 box. The type 1 box contains an American limited
edition keycard.

There is a closed American limited edition gate leading east. There is a closed
door leading south.

>
-= Scullery =-0/1

WARNING: your terminal doesn't support cursor position requests (CPR).
> take American limited edition keycard

You take the American limited edition keycard from the type 1 box.



>
-= Scullery =-0/2
```

```
nag186@login1.crc.pitt.edu ssh ~/project/text_world
tw-play tw_games/custom_game.z8
You begin to take stock of what's in the room.

You can make out a type 1 box. The type 1 box contains an American limited
edition keycard.

There is a closed American limited edition gate leading east. There is a closed
door leading south.

>
-= Scullery =-0/1

WARNING: your terminal doesn't support cursor position requests (CPR).
> take American limited edition keycard

You take the American limited edition keycard from the type 1 box.



>
-= Scullery =-0/2

> open American limited edition gate with American limited edition keycard

You unlock American limited edition gate.


>
-= Scullery =-0/3

>
```

Task

- Create a unified TextWorld benchmark for multiple game setups.
- Ensure strategy comparison under controlled, reproducible conditions.
- Analyze performance feedback per task and strategy.

Literature Review

- Research strategies for text-based games fall into four main categories:
  - LLM Prompting: Zhuo & Murata (2024),Phan et al. (2025)
  - Self-Reflection: Lippmann et al. (2024): Sweet & Sour reflection in ScienceWorld.
  - Toolchain and Memory: Yao et al. (2023),Zhang & Long (2025)
  - Reinforcement Learning: Gruppi et al. (2024)
- Each strategy focuses on different ways of enhancing LLM reasoning and planning.

Dataset

- TextWorld (Côté et al., 2019): procedural game generator for NLP tasks.
- Generates reproducible quests with varying difficulty and structure.
- Gold labels include valid action space and optimal action sequence.
- Enables standardized, objective model evaluation.

Data Description & Models

- Input: Text environment from TextWorld setups (navigation, cooking, treasure hunt).
- Output: Model-generated text actions to achieve goals.
- Strategies evaluated:
- • LLM Prompting — raw reasoning ability.
- • Self-Reflection — iterative improvement.
- • Toolchain & Memory — external state support.
- • Reinforcement Learning — fine-tuned policies.

Methodology

- TextWorld Benchmark
- LLM Models and System Development
  - Plain LLM Prompting
  - LLM Self-reflection
  - LLM Toolchain
  - Reinforcement Learning

Evaluation

- The scoring primarily judges the performance on 3 metrics:
  - 1. The score given by the game (adjusted by the Handicap)
  - 2. The number of steps taken to complete the game
  - 3. The maximum handicap (the highest category of hint taken by the agent)

Category	Example Parameters/Values	Description
Game Type	Simple, Cooking, treasure hunter	Sets the type of goals and the task at hand
Grammar Settings	Theme, include adjectives, ambiguous instructions	Changes the difficulty in terms of language
Quest Scale Settings	Parallel quests, quest length, quest breadth	Changes the scale of the quest
Difficulty Settings	Level, rewards, recipe	Changes the difficulty in terms of the game itself

Ethics

- Embodied AI
  - Harmful effects of AI trained on games
  - Impact on job markets
- AI Personas
  - Methods inducing or amplifying biases against a race, gender, religion, etc.
  - Harmful language being used (abuses, racial slurs, etc.)

Steps

- TextWorld Benchmark Development
- LLM Models and System Development
  - Plain LLM Prompting
  - LLM Self-reflection
  - LLM Toolchain
  - Reinforcement Learning
- Evaluation of Techniques on the Benchmark
- Result Analysis

Roles

- Naman Gupta: Responsible for TextWorld Benchmark and Reinforcement Learning model development.
- Raul Viteri: Responsible for Self-reflection and Toolchain model development.
- Shaojun Zheng: Responsible for Plain LLM Prompting and Toolchain model development.
- All three group members are responsible for the evaluation and result analysis.

2. Charitha, Maria, Surabhi, Shubham, Victor, John

PROJECT
PROPOSAL

Classifying Multilingual Adversarial Prompts

Presented By

CHARITHA BATTINI | MARIA KYREZI | JOHN RAFLA | SURABHI RAGHAVAN | SHUBHAM SARVANKAR | VICTOR YU

A BRIEF **INTRODUCTION**

Classifying Multilingual Adversarial Prompts

Goal: Building a classifier that leverages a large, multilingual dataset to identify and distinguish between adversarial prompts across several categories.

Challenges:

- Existing datasets similar to ours do not match our scale
- Matching benchmarks set by classifiers trained on English-only datasets.

Key Impacts:

- Examines performance of various models/LLMs on under-represented languages.
- Curates a larger dataset that can be used to empower future work.
- Reduces under/over-blocking in cultural contexts.

LITERATURE **REVIEW**

What Past Studies Have Found

Cultural Awareness in LLMs: Inherent bias in LLMs trained in English introduce need for better differentiation between languages in red-teaming classification

Cross-lingual Safety Alignment: Most models produce unsafe outputs in non-English, low recourse languages, benchmarks show

Efforts Made to Increase Safety:

- *MultiJail* introduced vulnerabilities of low recourse languages
- *WildTeaming* expanded to 260K dataset, still English-centric
- *JailbreakBench* had more specific, multilingual category labels

Multilingual Jailbreaking Techniques:
Rendering *adversarial prompts as images* and multilingual *code-switching* found to increase successful attack rates

Detection Methods and Safety Classifiers:
Keyword based approaches fail for multilingual prompts, modern classifiers rely on multilingual embedding ⟶ better results using an mBERT or XLM-R model, also broader categorization

THE DATASET

Base Dataset: HarmBench Behaviors dataset

Languages: Greek, French, Arabic

Semantic Categories

1. Chemical-biological
2. Illegal
3. Misinformation_Disinformation
4. Harmful
5. Harassment_bullying
6. Cybercrime_intrusion
7. Copyright

Dataset size

- Initially, 1,530 harmful queries
- Translated into 3 languages
- A total of 6,120 queries

Translation Method

- State-of-the-art LLM for automatic translation
- Human annotation to ensure accuracy and prevent hallucinations

APPROACH & **METHODS**

XLM-RoBERTa: a transformer-based encoder trained on 2.5 TB of CommonCrawl text across 100 languages, making it particularly suited for our target languages (Greek, Arabic, French)

1. Data collection & preparation

- Gather multilingual data to build on the existing dataset
- Verify and sanitize data to ensure accuracy and ethical concerns

2. Setup baseline model

- TF-IDF + Logical Regression
- Character n-gram SVM
- Naive Bayes Classifier

3. Encoder-based model development

- Fine-tune XLM-RoBERTa model
- Extract sentence-level embedding for linguistic and semantic analysis

4. Decoder-based model evaluation

- Apply zero- and few-shot prompting using GPT-4o or mT5-XXL
- Experiment with direct and chain-of-thought prompting strategies
- Evaluate performance on consistent test sets for comparability

5. Adversial and cross-lingual testing

- Conduct robustness tests with adversarially modified prompts
- Compare encoder vs. decoder model performance under adversarial conditions

6. Evaluation and analysis

EVALUATION **FRAMEWORK** & **METRICS**

Our evaluation is designed to test accuracy, generalization, and robustness

- **Accuracy:** Can the model correctly classify unsafe prompts by harm type?
- **Generalization:** Does performance hold across languages, including unseen ones?
- **Robustness:** Does the model remain consistent under obfuscated or code-mixed inputs?

Aspect	Focus
Baselines	Keyword Heuristic, BERT (English-only), **XLM-RoBERTa** (ours), GPT-4o/mT5
Metrics	Accuracy, Precision, Recall, Macro-F1, CLGS (Cross-Lingual Generalization Score), ARR (Adversarial Robustness Rate)
Testing Setup	Zero-shot (held-out languages), joint multilingual, adversarial variants (ciphered, base64, code-mixed), optional human check
Success Targets	≥ 85 % Macro-F1, ≤ 10 % drop in zero-shot, ≥ 80 % robustness consistency

3. Uma, Yudan, Chase
Discourse Relation Classification

Task: Discourse Relation Classification

Goal: Create a system that works across multiple languages and annotation frameworks to identify connections between phrases in a text

- Connections can be :
  - Explicit: Evident through words such as "but", "therefore", or "because"
  - Implicit: Require contextual clues usually from surrounding text
- Challenges:
  - Dependencies on a lot of text to make a connection
  - Scarce data for uncommon languages

Input and Output

Input: 2 distinct text arguments

Output: A specific discourse relation label from a predefined set

Example:

- Explicit Relation:
  - Input:
    - Arg1: "The class was very difficult.:
    - Arg2: "**Therefore,** many students failed the exam."
  - Output: Cause.Result
- Implicit Relation:
  - Input:
    - Arg1: "The class was very difficult."
    - Arg2: "Many students failed the exam."
  - Output: Cause.Result

Literature Review

1. (Dai and Huang, 2018)
    - Focused on feature engineering, such as using paragraph-level context, to improve the performance of custom neural models.
2. (Eichin et al., 2025)
    - Shows that large, multilingual LLMs can identify discourse relations without task-specific fine-tuning, suggesting they learn these patterns during pre-training.
3. (Ju et al., 2025)
    - Shows other modern approaches and uses strategies such as translating text to help low-resource languages (data augmentation) to achieve results.

Data

Dataset: Multilingual and multi-framework dataset from the 2025 Discourse Relation Parsing and Treebanking (DISRPT) shared task

- Includes numerous languages such as English, Czech, Persian, Thai, etc.
- Annotated with labels under established frameworks
- Some datasets are under licensing restrictions from the Linguistic Daat Consortium, but we are planning to request access through the Pitt library

Methods

3 primary modelling approaches (using models with < 4 billion parameters)

- Encoder-Only Model: Fine-tune an encoder model like ModernBERT for the classification task.
- Decoder-Only Model (Instruction Tuning): Use instruction-style fine-tuning on a decoder model like QWEN 3, Llama 3.2 3B, or Gemma 3.
- Decoder-Only Model (RL-style Tuning): Experiment with reinforcement learning-based fine-tuning, possibly using GRPO.
- *Plan to incorporate feature engineering and fine-tuning

Evaluation

Performance Metrics:

- Accuracy, Precision, Recall, F1-Score

Evaluation Script:

- Taken from the shared task

Comparison Baselines:

- Published results from DeDisCo
- Classical stats methods such as n-grams or a base BERT model

Ethics

- Performance Disparity
  - Unexpected performance gap in high and low resource languages
  - Could lead to inequities in things like text summarization or machine translations
- Data Augmentation Bias
  - Machine-translations can contain errors and cannot usually capture nuances of languages

Steps

1. Dataset Access
2. Data Pre-processing
3. Data Augmentation
4. Modeling
5. Computing Environment
6. Model Evaluation
7. Error Analysis
8. Reporting

4. Nate, Zhiwei, Hongbo

AI vs. Human: Binary Detection of Generated Text

Motivation

- **Binary Machine-Generated Text Detection**
- **The Challenge**
  - Generative AI produces "human-like" text.
  - This threatens the integrity and credibility of online information.
  - Detection is difficult due to model diversity, domain variation, and language differences.
- **Our Goal**
  - **Task:** Binary Classification (AI vs. Human).
  - **Input:** A short text passage.
  - **Output:**
    - **Label 0:** Human-written
    - **Label 1:** Machine-generated
  - To reproduce and achieve improvement over the official methods.

Related Works

- **Single Models** (e.g., LSTM, BERT-CNN)
  - Show high accuracy (97%-99%) in specific contexts.
  - May not generalize well (e.g., trained on older models like ChatGPT-2).
- **Ensemble Methods** (e.g., DeBERTa + RoBERTa)
  - Combining models improves accuracy over single-model performance.
  - Demonstrates the value of multiple perspectives.
- **NLP Metrics** (Perplexity & Burstiness)
  - Provide small but consistent accuracy improvements (~1%).
  - **Key Insight**: Linguistic features are a complementary signal to neural models.

Dataset & Method Overview

- **Dataset**
  - Official COLING 2025 Shared Task 1 data. [URL: Shared Tasks]
  - Multi-domain, multi-lingual mix of human and AI text.
- **Method: A Hybrid Detection Framework**
  - **Core:** Centered on the Qwen3-8B model.
  - **Ensemble:** Combines three detectors via Soft Voting.
- **Three Complementary Detectors:**
  - **Qwen3-8B (PEFT):** For contextual classification.
  - **Mamba (Perplexity):** For fluency/fluctuation patterns.
  - **RoBERTa (Encoder):** For semantic/syntactic cues.

Methodology Deep Dive

- **1. Qwen3-8B (PEFT)**
  - Uses Parameter-Efficient Fine-Tuning (e.g., LoRA).
  - Efficiently adapts the large model for binary classification.
- **2. Mamba (Perplexity Sequence)**
  - Step 1: Qwen3-8B generates a token-level perplexity sequence.
  - Step 2: Mamba model processes this sequence to detect fluency patterns.
- **3. RoBERTa Encoder**
  - A fine-tuned encoder (e.g., RoBERTa).
  - Captures complementary semantic and syntactic cues.
- **Final Ensemble**
  - A weighted average of the three calibrated probabilities.
  - Weights are tuned on the dev set to maximize Micro-F1

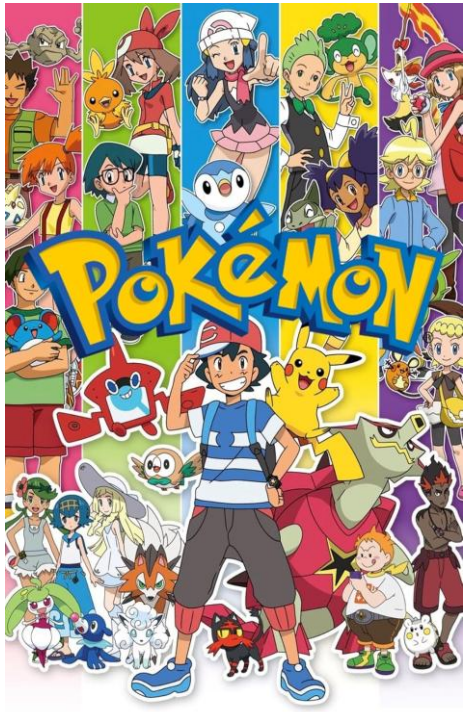
Evaluation

- **Primary Metric: Micro-F1 Score**
  - The official metric for the task.
  - Balances precision and recall.
  - Reliable for (potentially) imbalanced datasets.
- **Secondary Metrics**
  - Accuracy
  - Precision
  - Recall
- **Process**
  - All models evaluated on the official dev/test splits using the provided scripts.
- **Challenges**
  - Achieve each method effectively.
  - Find a best combination of the weight for each method on dev set while prevent overfit.

4. Tim, Keshav, Lucy

Pokemon!!!


MANTYKE Lv23
LUCARIO Lv30
88/88
What will LUCARIO do?


Pokémon


STAGE 2
Evolves from Charmeleon Put Charizard on the Stage 1 card
Charizard
120 HP
Flame Pokémon. Length: 5' 7", Weight: 199.5 lbs.
Pokémon Power: Energy Burn All Energy attached to this Pokémon are Energy instead of their usual type. This Pokémon Power stops working if this Pokémon is Asleep, Confused, or Paralyzed.
Fire Spin Discard 2 Energy from this Pokémon. 100
weakness
resistance -30
retreat cost
Spits fire that is hot enough to melt boulders. Known to unintentionally cause forest fires. LV. 76 #6
003/034
Illus. Mitsuhiro Arita
©2023 Pokémon / Nintendo / Creatures / GAME FREAK

Previous work

- Mika Hämäläinen, Khalid Alnajjar, and Niko Partanen. 2021. How cute is pikachu? gathering and ranking pokémon properties from data with pokémon word embeddings. Preprint, arXiv:2108.09546.
- Laura Cabello, Jiaang Li, and Ilias Chalkidis. 2023. Pokemonchat: Auditing chatgpt for pokémon universe knowledge. Preprint, arXiv:2306.03024.
- Tadisetty Sai Yashwanth and Dhatri C. 2025. A multi- agent pokemon tournament for evaluating strate- gic reasoning of large language models. Preprint, arXiv:2508.01623.
- Ryan Belfer. 2021. Predicting pokémon type using the pokédex. https://medium. com/analytics-vidhya/predicting-pok% C3%A9mon-type-with-the-pok%C3% A9dex-7038754dc422. Accessed: 2025-10-13.
- Shigeto Kawahara, Atsushi Noto, and Gakuji Kumagai. 2018. Sound symbolic patterns in pokémon names. Phonetica, 75(3):219–244

Bulbapedia

Biology

Furret is a long, slim-bodied mustelid Pokémon with cream-colored fur and dark brown rings along the length of its body. The back of its head and neck are also dark brown; this marking extends to below its arms where the first ring is formed. It has cream-colored tips on its ears, two brown, whisker-like markings on each cheek, and round, black eyes. Its four limbs are stubby; the forepaws are brown and the hind paws are cream-colored. Furret is capable of standing on its hind legs but prefers to move on all fours. Its body and tail are so similar in structure that it is impossible to tell where its tail begins.

Despite its short limbs, Furret is very quick and agile. This allows Furret to escape even when in the arms of a human. Its speed allows it to catch prey such as Rattata. It tends to burrow under the ground of meadows and other temperate grasslands. Its narrow burrows are well-suited for its slim body and are very difficult for other Pokémon to enter. The burrows also become more maze-like deeper inside, which makes it even harder to find Furret's nest. A mother Furret curls itself around its offspring to help them sleep. Once the offspring are older, the mother Furret will take them outside to prepare them for independence. In Lental, Mightyena is a natural predator of Furret.

Type	Normal
Abilities	Run Away or Keen Eye; Frisk (Hidden Ability)
Gender ratio	50% male, 50% female
Catch rate	90 (20.1%)
Breeding	
Egg Group	Field
Hatch time	15 cycles
Height	5'11" 1.8 m
Weight	71.7 lbs. 32.5 kg
Base experience yield	116 Gen. II-IV; 145 V+
Leveling rate	Medium Fast
EV yield	Total: 2
	0 HP; 0 Atk; 0 Def; 0 Sp.Atk; 0 Sp.Def; 2 Speed
Shape	
Footprint	
Pokédex color	Brown
Base friendship	70

Pokedex

Generation VI — Kalos Central #110 | Hoenn #—

X	The mother puts its offspring to sleep by curling up around them. It corners foes with speed.
Y	It makes a nest to suit its long and skinny body. The nest is impossible for other Pokémon to enter.
Omega Ruby / Alpha Sapphire	Furret has a very slim build. When under attack, it can slickly squirm through narrow spaces and get away. In spite of its short limbs, this Pokémon is very nimble and fleet.

Generation VII — Alola #— | Kanto #—

This Pokémon has no Pokédex entries in Generation VII.

Generation VIII — Galar #— | Sinnoh #— | Hisui #—

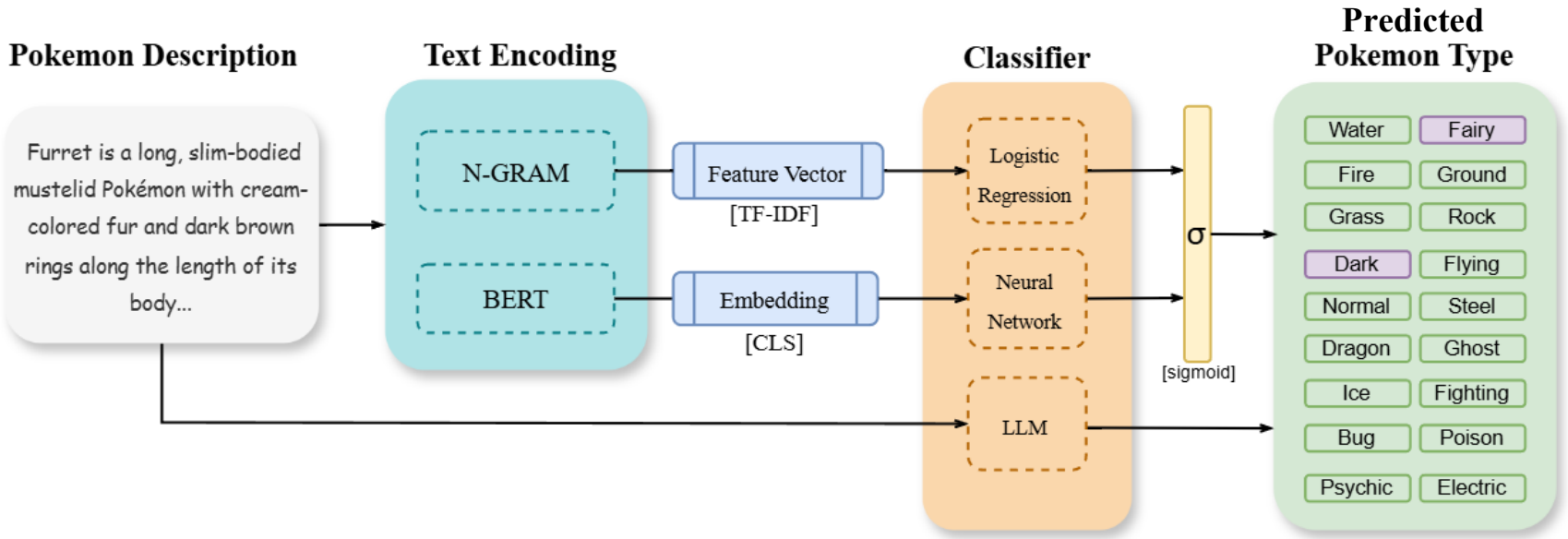
This Pokémon has no Pokédex entries in Sword, Shield, and Legends: Arceus.

Brilliant Diamond / Shining Pearl	The mother puts its offspring to sleep by curling up around them. It corners foes with speed.

Generation IX — Paldea #— | Kitakami #27 | Blueberry #—

Scarlet	It raises its offspring inside a long, narrow nest. Once they're old enough, it takes them outside the nest to prepare them for independence.
Violet	It is nimble and has a very flexible body. Even if you get ahold of it, it'll slip right out of your arms.

What are we doing?



Ground truth: [water:0, Fairy: 0, normal: 1 ….]

[water: 0.001, Fairy: .52, Normal: .02...]

Given the following description, select up to two types that best classify the Pokemon.
"The mother puts its offspring to sleep by curling up around them. It corners foes with speed...."

possible types: [Grass, Water, Fire, Electric, Ice, Fighting, Poison, Dark, Fairy, Steel, Flying, Normal, Psychic,
Ghost, Ground, Rock, Dragon, Bug]

Provide the output in the following structure: [type]

[normal]

Evaluation

Metrics:

- F1 scores:
  - micro: overall performance
  - macro: since the dataset is imbalanced
  - Confusion matrix

Challenges:

- Small dataset:
- 1025 examples with rare labels
- iterative stratification (preserve single type freq + combos across CV folds)

- start with: 80/10/10 split and go from there