11-891 Report 1: Dataset Analysis and Baseline/Method Proposal

First Last 1* **First Last 2*** **First Last 3*** **First Last 4***
{ID1, ID2, ID3, ID4}@andrew.cmu.edu

1 Task Definition and Dataset Choice (1 page)

1.1 What phenomena or task will you work on?

1.2 What about this phenomena or task fundamentally involves code?

1.3 Does a dataset exist, or will you be collecting it, or does your task not require a dataset?

If the dataset exists already, reference it here. If it doesn't, detail how you'll collect it and how long it will take, or justify why your task doesn't require a dataset.

1.4 Expertise

We have the following expertise in the underlying aspects required by this project

1. Team member 1: Research paper in static analysis, ...
2. Team member 2: Took NLP in Fall 2023, ...
3. ...

*Everyone Contributed Equally – Alphabetical order

2 Dataset Analysis (1 page)

2.1 Dataset properties

(GBs, code languages, numbers of examples, ...)

2.2 Instance analysis

(use a small sample – e.g. validation splits):

1. Code diversity: e.g. syntactic complexity, API usage, input and output spaces

2. If your task/dataset involves language: Lexical diversity, sentence length, ...

2.3 Evaluation metrics

What intrinsic (intermediate) and extrinsic (end-task-related) evaluations will you use to know if things are working?

3 Related Work and Background

Aim for 5 papers per person, which can go into the categories below.

Related Datasets

Baselines

Prior Work

Relevant Techniques

4 Baselines and Proposed Approach (1 page)

4.1 Baselines

Ideally at least two baselines, although one is ok if it will be difficult to implement (e.g. no publicly released code or models).

We will ask you to have results from these, and analysis of them, in Report 2.

4.2 Proposed Approach(es)

What might you do beyond the baseline?

4.3 Compute requirements

1. Files (if you have a large training dataset – can fit in RAM?)
2. Models (can fit on GCP/AWS GPUs?)
3. Training (if you're training models, estimated GPU hours)
4. Inference (how many instances in your task/dataset? Do you need to sample many outputs? Execute code?)

5 Team member contributions

Member 1 contributed ...

Member 2 contributed ...

... contributed ...

References