# Dynamic-SUPERB Tutorial

Chien-yu Huang
Speech Processing Lab., National Taiwan University

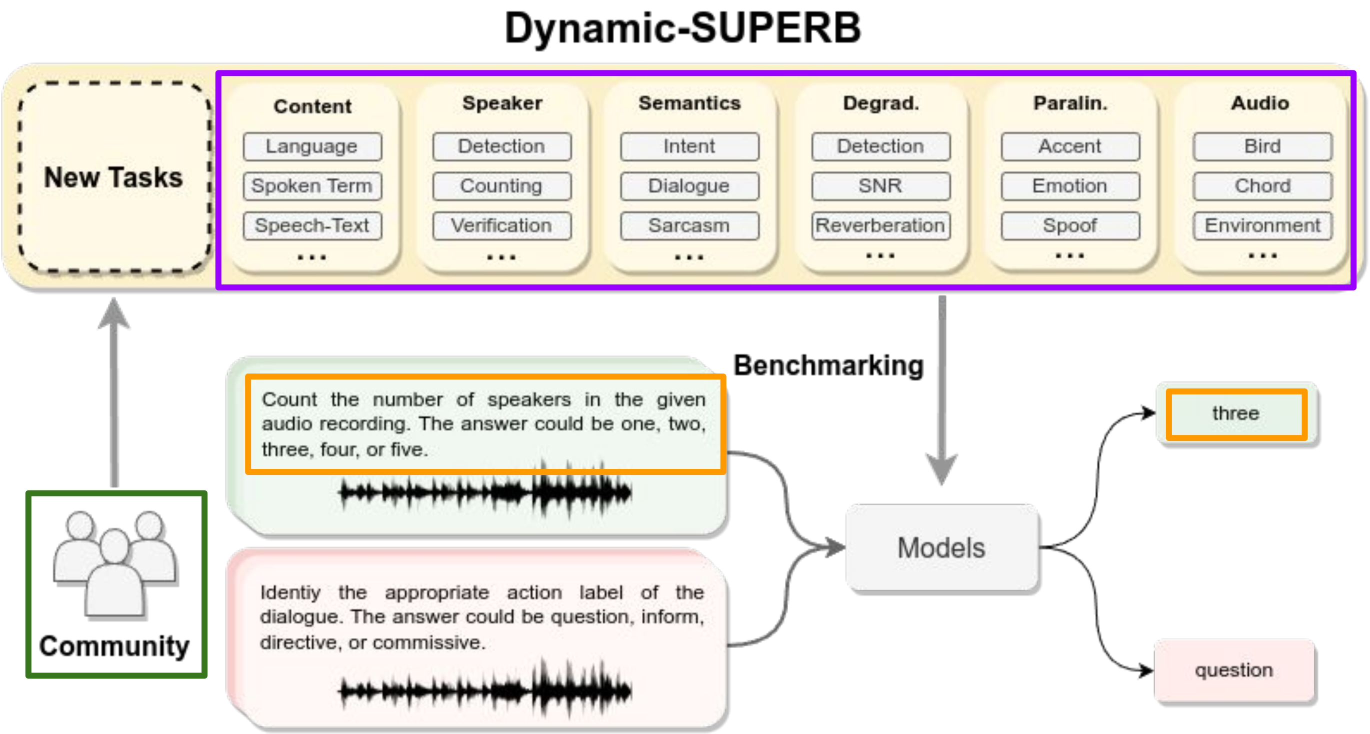
# Overview

- Framework overview
- Benchmark tasks
- Evaluations
- Baseline models
- Score submission
- Task contribution

# Overview

- Framework overview
- Benchmark tasks
- Evaluations
- Baseline models
- Score submission
- Task contribution

# Framework Overview



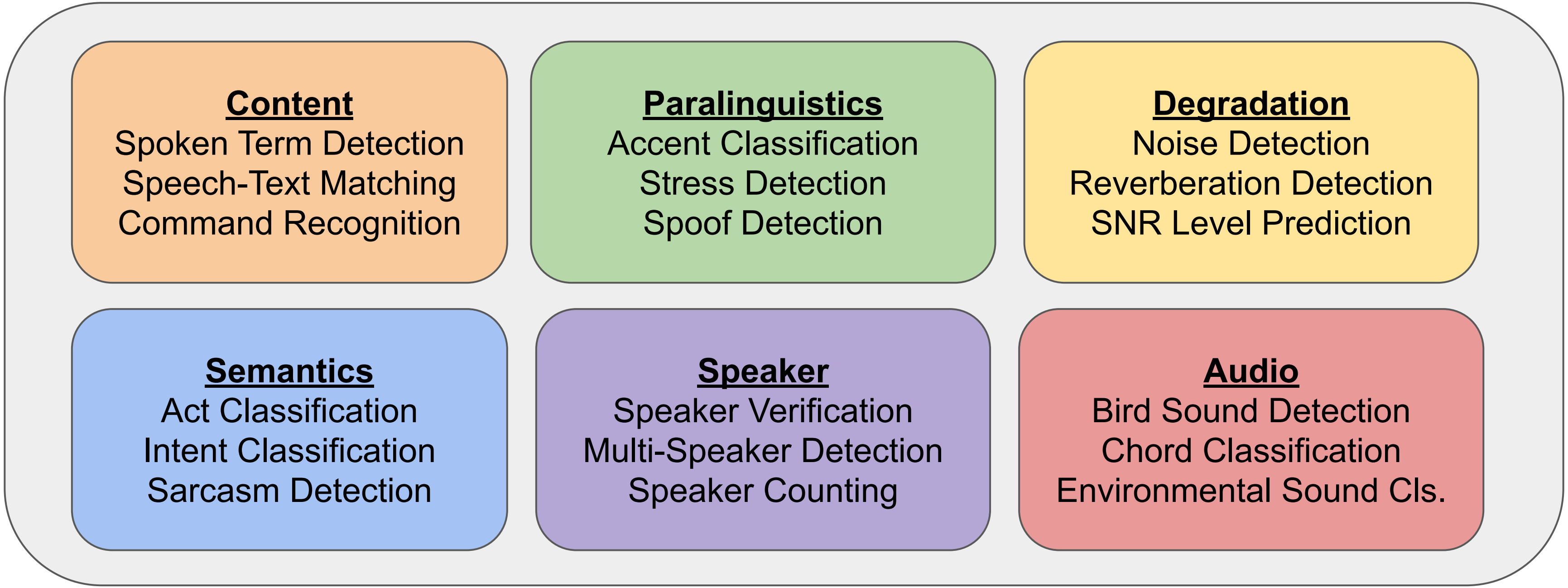
Count the number of speakers in the given audio recording. The answer could be one, two, three, four, or five.

Identiy the appropriate action label of the dialogue. The answer could be question, inform, directive, or commissive.

# Overview

- Framework overview
- Benchmark tasks
- Evaluations
- Baseline models
- Score submission
- Task contribution

# Benchmark Tasks



**Content**
Spoken Term Detection
Speech-Text Matching
Command Recognition

**Paralinguistics**
Accent Classification
Stress Detection
Spoof Detection

**Degradation**
Noise Detection
Reverberation Detection
SNR Level Prediction

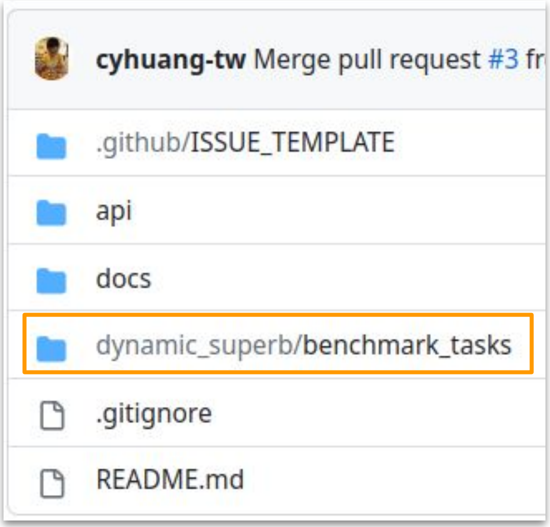
**Semantics**
Act Classification
Intent Classification
Sarcasm Detection

**Speaker**
Speaker Verification
Multi-Speaker Detection
Speaker Counting

**Audio**
Bird Sound Detection
Chord Classification
Environmental Sound Cls.

- Covers 6 dimensions, 33 tasks, and 55 evaluation instances.
- They are all classification tasks.

# Task Format



**cyhuang-tw** Merge pull request #3 fr

.github/ISSUE_TEMPLATE

api

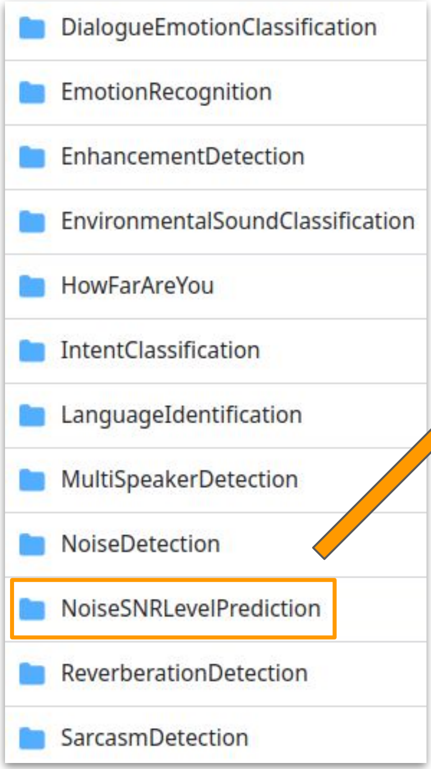
docs

dynamic_superb/benchmark_tasks

.gitignore

README.md

## Tasks



DialogueEmotionClassification

EmotionRecognition

EnhancementDetection

EnvironmentalSoundClassification

HowFarAreYou

IntentClassification

LanguageIdentification

MultiSpeakerDetection

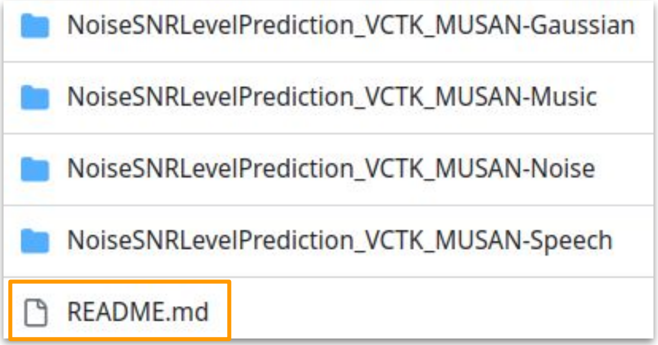
NoiseDetection

NoiseSNRLevelPrediction

ReverberationDetection

SarcasmDetection

## Evaluation Instances



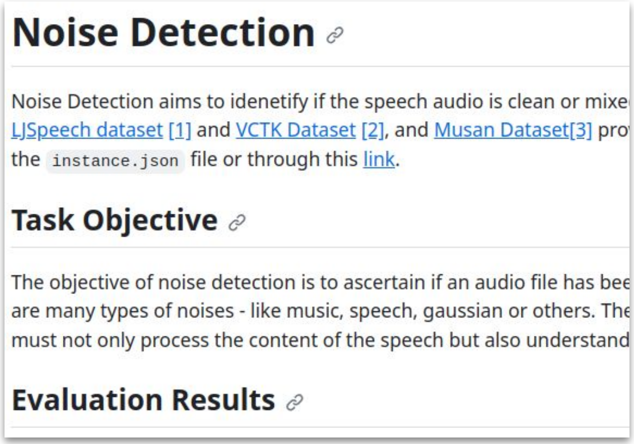
NoiseSNRLevelPrediction_VCTK_MUSAN-Gaussian

NoiseSNRLevelPrediction_VCTK_MUSAN-Music

NoiseSNRLevelPrediction_VCTK_MUSAN-Noise

NoiseSNRLevelPrediction_VCTK_MUSAN-Speech

README.md



# Noise Detection

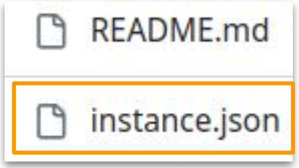
Noise Detection aims to idenetify if the speech audio is clean or mixe
LJSpeech dataset [1] and VCTK Dataset [2], and Musan Dataset[3] pro
the `instance.json` file or through this link.

## Task Objective

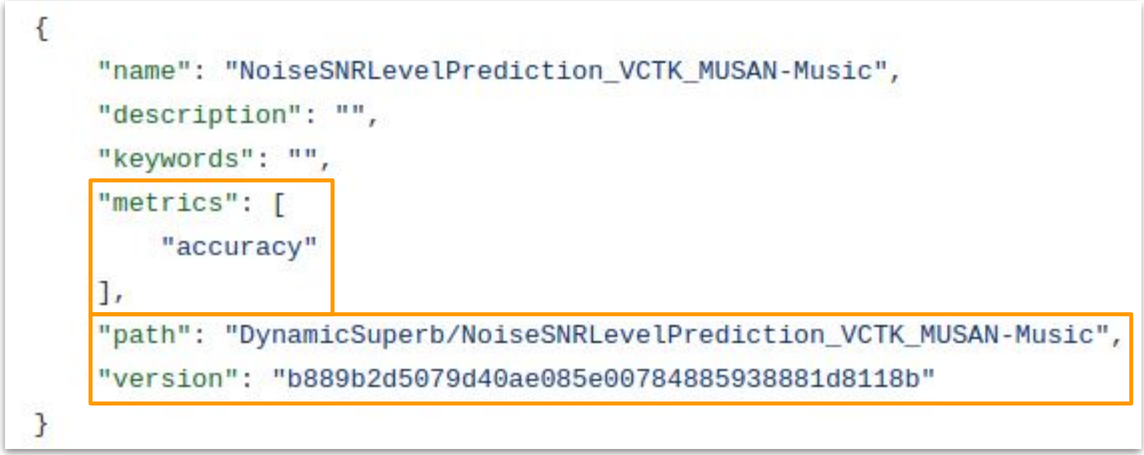
The objective of noise detection is to ascertain if an audio file has bee
are many types of noises - like music, speech, gaussian or others. The
must not only process the content of the speech but also understand

## Evaluation Results

# Task Format



README.md
instance.json



```
{
    "name": "NoiseSNRLevelPrediction_VCTK_MUSAN-Music",
    "description": "",
    "keywords": "",
    "metrics": [
        "accuracy"
    ],
    "path": "DynamicSuperb/NoiseSNRLevelPrediction_VCTK_MUSAN-Music",
    "version": "b889b2d5079d40ae085e00784885938881d8118b"
}
```

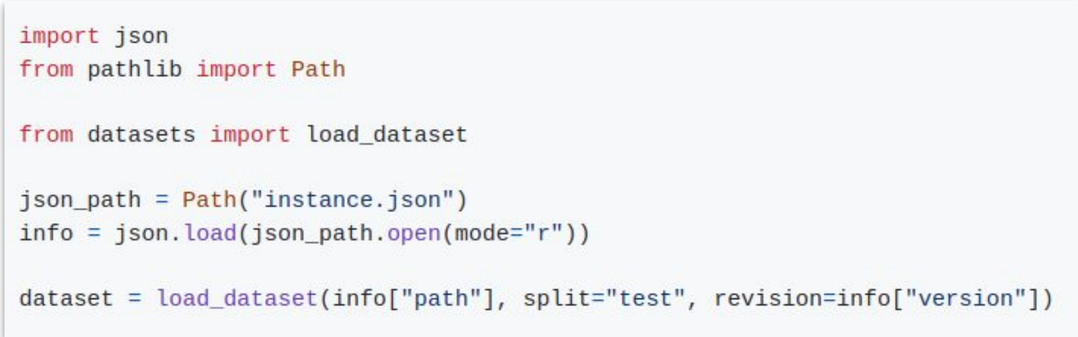
- Access all information in “instance.json”.
- Download data with “path” and “version” from Huggingface.

# Overview

- Framework overview
- Benchmark tasks
- Evaluations
- Baseline models
- Score submission
- Task contribution

# Evaluation - Integrating Huggingface API

- Download data from Huggingface with datasets package (no explicit download).



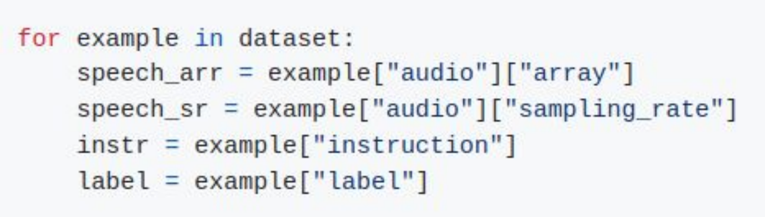
```
import json
from pathlib import Path

from datasets import load_dataset

json_path = Path("instance.json")
info = json.load(json_path.open(mode="r"))

dataset = load_dataset(info["path"], split="test", revision=info["version"])
```

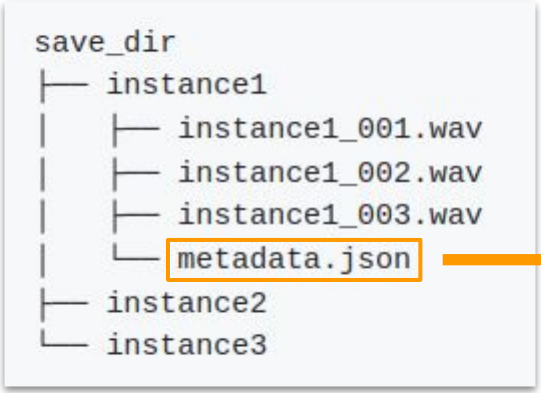
- Iterate with a very simple for loop.



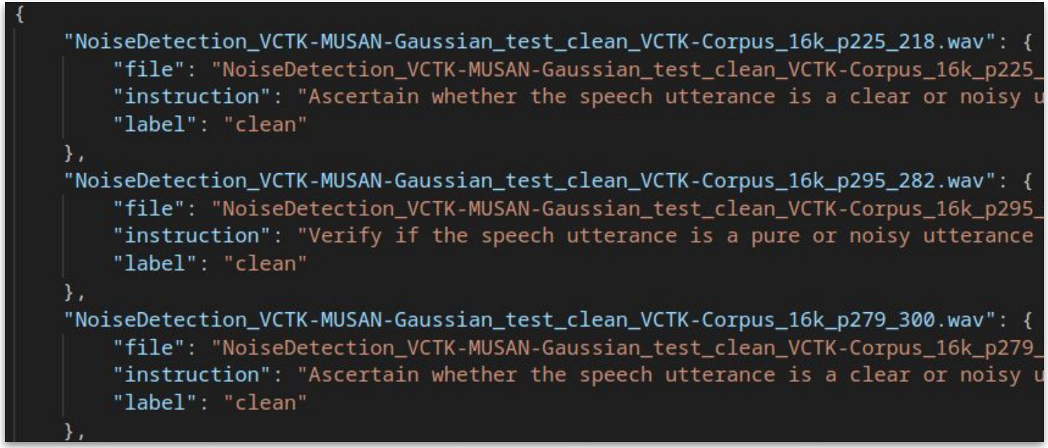
```
for example in dataset:
    speech_arr = example["audio"]["array"]
    speech_sr = example["audio"]["sampling_rate"]
    instr = example["instruction"]
    label = example["label"]
```

# Evaluation - Save to Local Files

- Save files to local storage explicitly for easy modification.
- Scripts in api/preprocess.



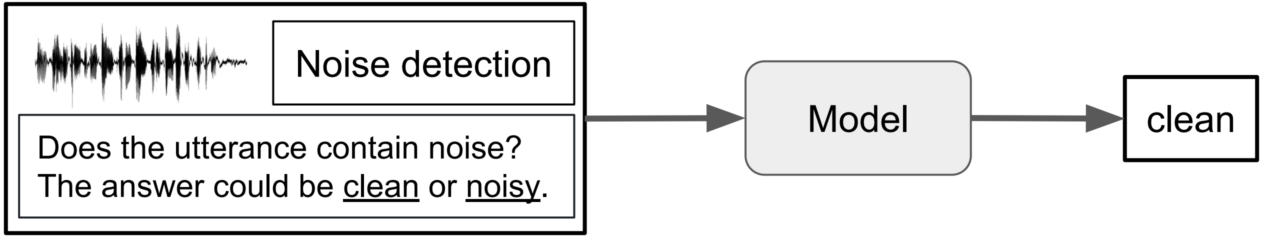
```
save_dir
├── instance1
│    ├── instance1_001.wav
│    ├── instance1_002.wav
│    ├── instance1_003.wav
│    └── metadata.json
├── instance2
└── instance3
```



```
{
    "NoiseDetection_VCTK-MUSAN-Gaussian_test_clean_VCTK-Corpus_16k_p225_218.wav": {
        "file": "NoiseDetection_VCTK-MUSAN-Gaussian_test_clean_VCTK-Corpus_16k_p225_
        "instruction": "Ascertain whether the speech utterance is a clear or noisy u
        "label": "clean"
    },
    "NoiseDetection_VCTK-MUSAN-Gaussian_test_clean_VCTK-Corpus_16k_p295_282.wav": {
        "file": "NoiseDetection_VCTK-MUSAN-Gaussian_test_clean_VCTK-Corpus_16k_p295_
        "instruction": "Verify if the speech utterance is a pure or noisy utterance
        "label": "clean"
    },
    "NoiseDetection_VCTK-MUSAN-Gaussian_test_clean_VCTK-Corpus_16k_p279_300.wav": {
        "file": "NoiseDetection_VCTK-MUSAN-Gaussian_test_clean_VCTK-Corpus_16k_p279_
        "instruction": "Ascertain whether the speech utterance is a clear or noisy u
        "label": "clean"
    },
```

# Evaluation - Metrics

- Accuracy for classification tasks (string matching, case-insensitive).



Noise detection

Does the utterance contain noise?
The answer could be clean or noisy.

Groundtruth: clean

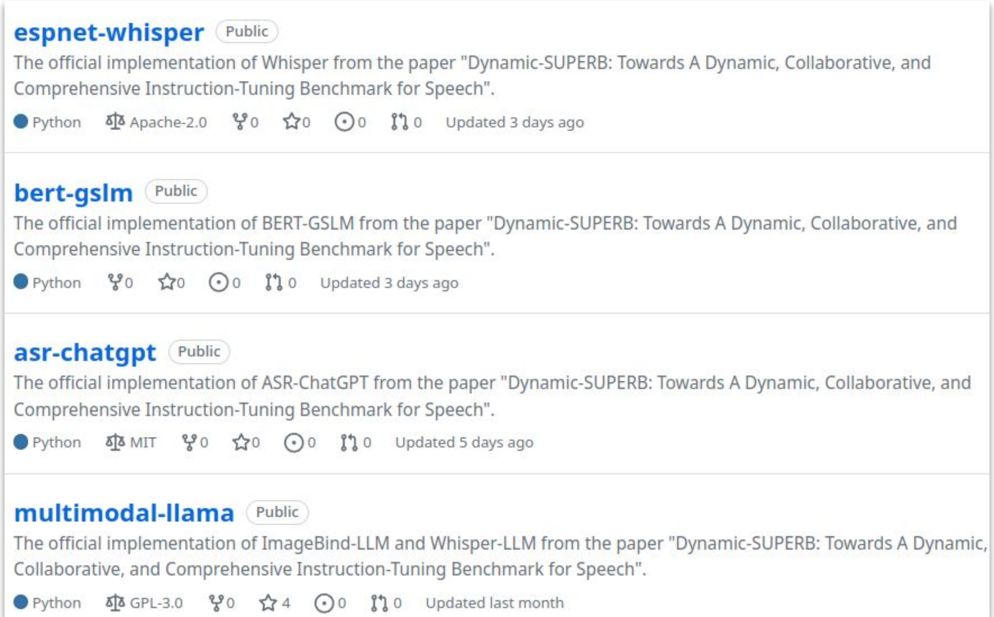
|  |  | case-insensitive | redundant comma | synonyms |
|---|---|---|---|---|
| **Prediction** | clean | Clean | clean, | clear |
| **Matched** | ✓ | ✓ | ✘ | ✘ |

- Not the best choice for free-form responses from LLMs.
- Working on more flexible measures (e.g., sentence embeddings).

# Overview

- Framework overview
- Benchmark tasks
- Evaluations
- Baseline models
- Score submission
- Task contribution

# Baseline Models - Implementations



**espnet-whisper** Public

The official implementation of Whisper from the paper "Dynamic-SUPERB: Towards A Dynamic, Collaborative, and Comprehensive Instruction-Tuning Benchmark for Speech".

● Python · Apache-2.0 · 0 · ☆0 · 0 · 0 · Updated 3 days ago

**bert-gslm** Public

The official implementation of BERT-GSLM from the paper "Dynamic-SUPERB: Towards A Dynamic, Collaborative, and Comprehensive Instruction-Tuning Benchmark for Speech".

● Python · 0 · ☆0 · 0 · 0 · Updated 3 days ago

**asr-chatgpt** Public

The official implementation of ASR-ChatGPT from the paper "Dynamic-SUPERB: Towards A Dynamic, Collaborative, and Comprehensive Instruction-Tuning Benchmark for Speech".

● Python · MIT · 0 · ☆0 · 0 · 0 · Updated 5 days ago

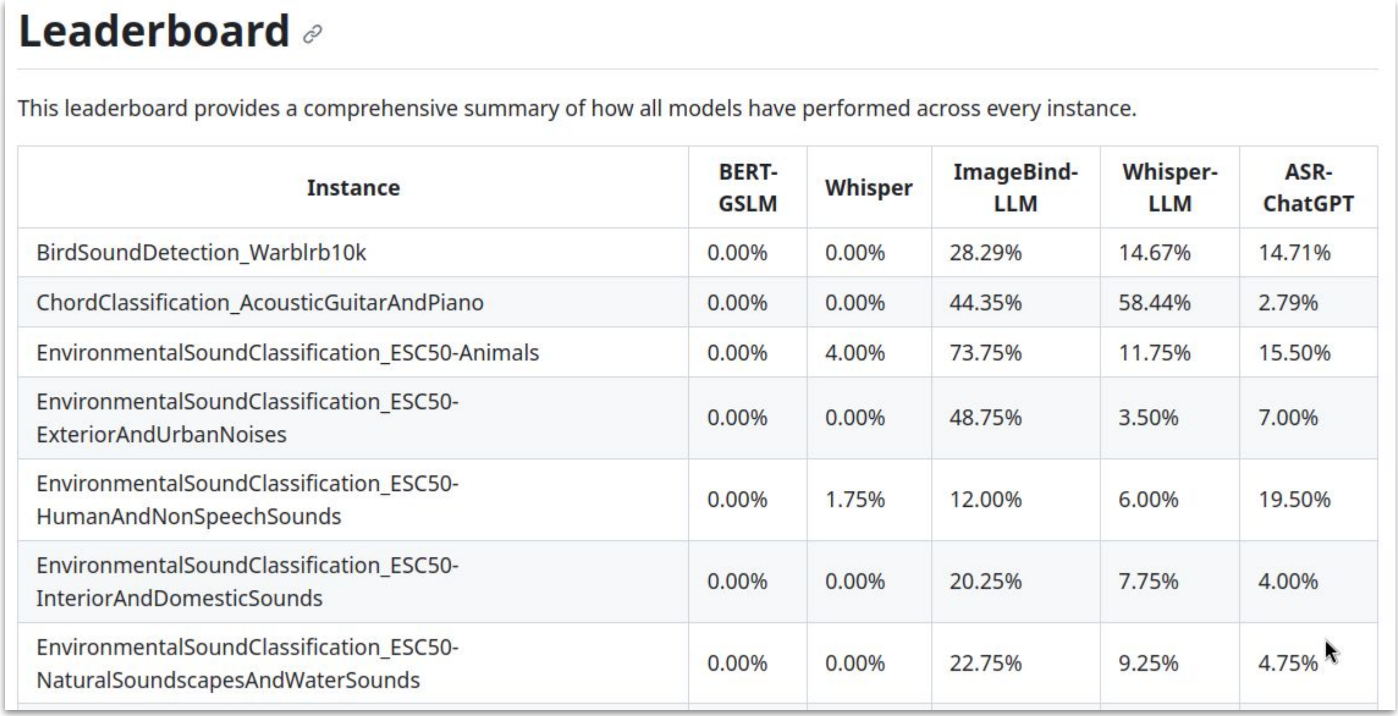
**multimodal-llama** Public

The official implementation of ImageBind-LLM and Whisper-LLM from the paper "Dynamic-SUPERB: Towards A Dynamic, Collaborative, and Comprehensive Instruction-Tuning Benchmark for Speech".

● Python · GPL-3.0 · 0 · ☆4 · 0 · 0 · Updated last month

- Open-sourced all baselines used in the Dynamic-SUPERB paper.
- Detailed guideline for running inference with pre-trained weights.

# Baseline Models - Performance



## Leaderboard

This leaderboard provides a comprehensive summary of how all models have performed across every instance.

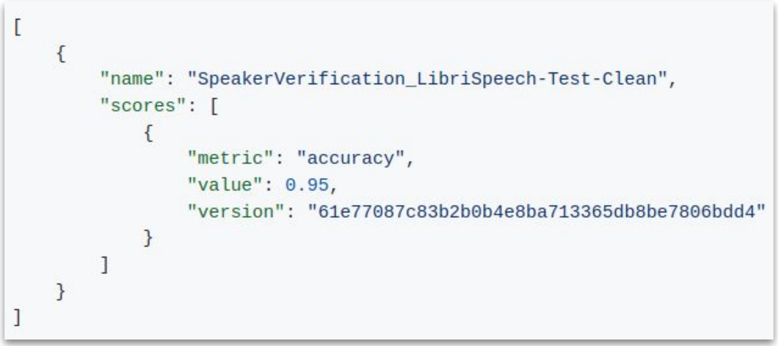
| Instance | BERT-GSLM | Whisper | ImageBind-LLM | Whisper-LLM | ASR-ChatGPT |
|---|---|---|---|---|---|
| BirdSoundDetection_Warblrb10k | 0.00% | 0.00% | 28.29% | 14.67% | 14.71% |
| ChordClassification_AcousticGuitarAndPiano | 0.00% | 0.00% | 44.35% | 58.44% | 2.79% |
| EnvironmentalSoundClassification_ESC50-Animals | 0.00% | 4.00% | 73.75% | 11.75% | 15.50% |
| EnvironmentalSoundClassification_ESC50-ExteriorAndUrbanNoises | 0.00% | 0.00% | 48.75% | 3.50% | 7.00% |
| EnvironmentalSoundClassification_ESC50-HumanAndNonSpeechSounds | 0.00% | 1.75% | 12.00% | 6.00% | 19.50% |
| EnvironmentalSoundClassification_ESC50-InteriorAndDomesticSounds | 0.00% | 0.00% | 20.25% | 7.75% | 4.00% |
| EnvironmentalSoundClassification_ESC50-NaturalSoundscapesAndWaterSounds | 0.00% | 0.00% | 22.75% | 9.25% | 4.75% |

# Overview

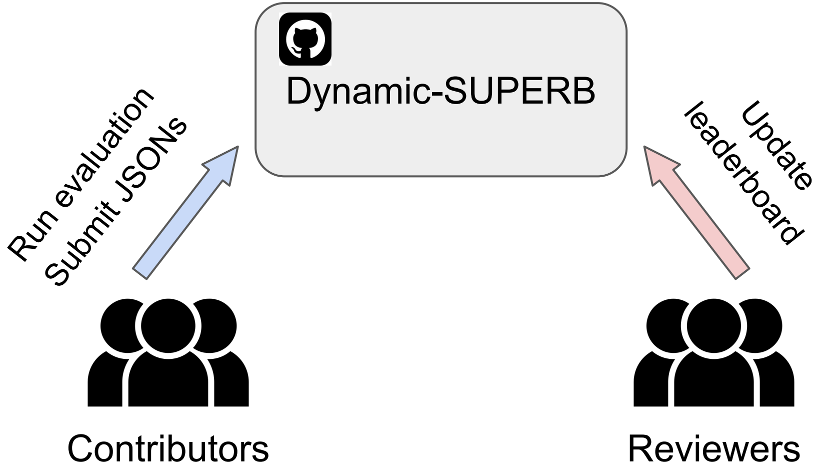
- Framework overview
- Benchmark tasks
- Evaluations
- Baseline models
- Score submission
- Task contribution

# Score Submission

- Pack all scores into a JSON file & submit via pull requests.
- No need to evaluate on all instances.



```
[
    {
        "name": "SpeakerVerification_LibriSpeech-Test-Clean",
        "scores": [
            {
                "metric": "accuracy",
                "value": 0.95,
                "version": "61e77087c83b2b0b4e8ba713365db8be7806bdd4"
            }
        ]
    }
]
```



Dynamic-SUPERB

Run evaluation
Submit JSONs

Update
leaderboard

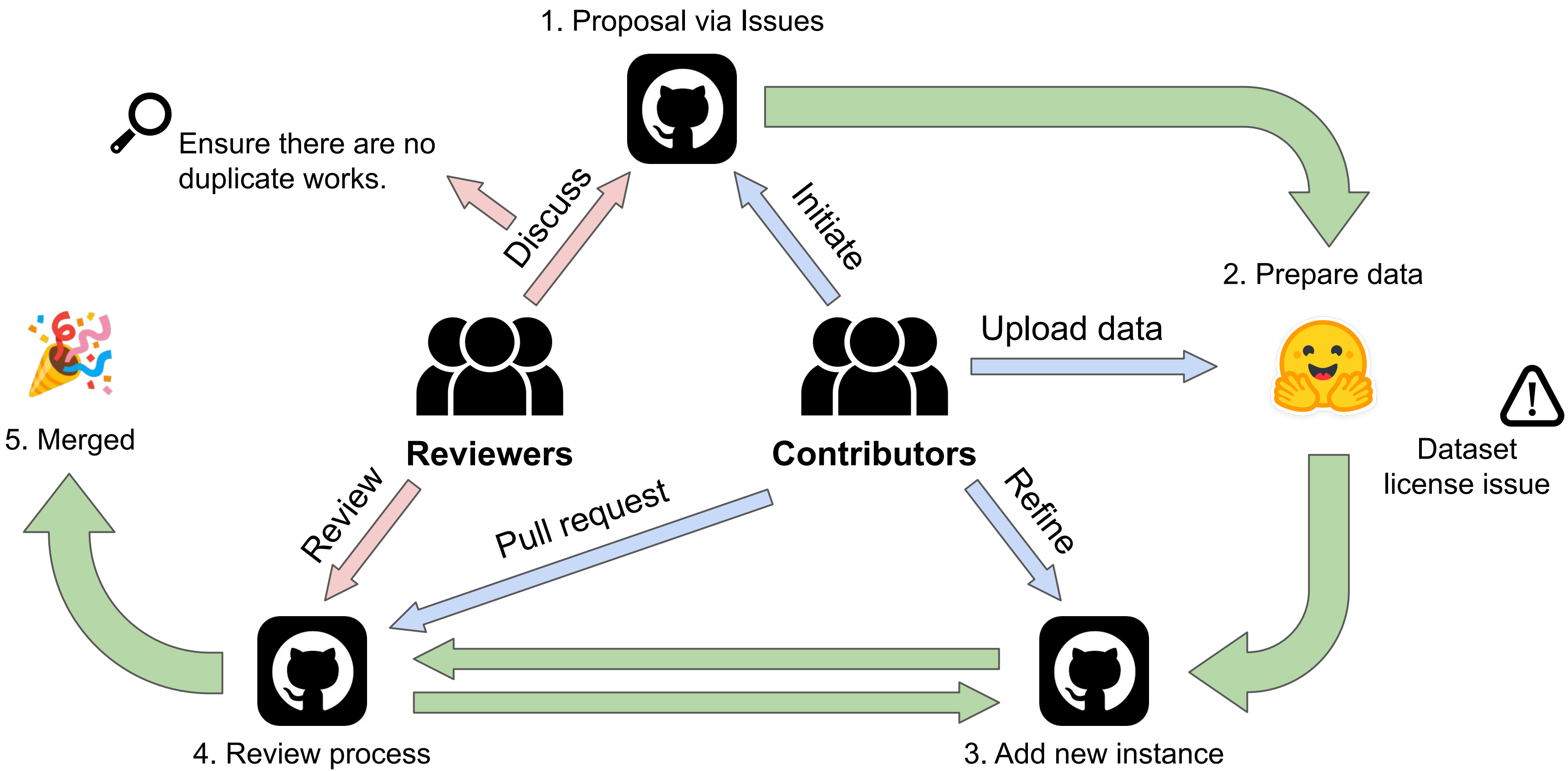
Contributors

Reviewers

# Overview

- Framework overview
- Benchmark tasks
- Evaluations
- Baseline models
- Score submission
- Task contribution

# Task Contribution



# Task Contribution

- Only single-label classification tasks in benchmark now.
- Generative tasks are also important.
  - Speech recognition
  - Text-to-speech
  - Voice conversion
  - Speech enhancement
  - Speech translation
- Welcome contributions on various metrics (also through PRs).