Dynamic Eval Contribution Guide

This document is meant to serve as both a way of organizing resources for the Dynamic Evaluation project, as well as to outline the many ways that it is possible to contribute to the project if you have any interest/availability!

There are currently several avenues for collaboration:

1. Adding Metrics and references to the TODO list
2. Adding Metrics to the Metric Bank
3. Adding Task recommendations to the TODO list
4. Adding Tasks to our Evaluation Suite
5. (Upcoming) Writing feedback

Relevant Resources for the Project:

Github: https://github.com/XenonMolecule/autometrics
Overleaf: https://www.overleaf.com/7857771127tbpzwgqbsfgx#8730b5
Weekly Slides: Slides
Metric Todo List: Metric ToDo
Task Todo List: Task ToDo
Random Notes/Brainstorming: Brainstorming and Notes

Adding Metrics to ToDo List:

Feel free to add whatever level of detail you have in mind to the sheet of raw metric data:
Metric ToDo

Metric name	Display Name	Metric Family	Measures	Framing/task	Trainable?	Uses source?	Uses ref?	# papers	Paper IDs	URLs
bleu	BLEU	BLEU	Overlap		FALSE	FALSE	TRUE	39	2023.inlg-main.2	https://aclanthology.org/2023.inlg-main.2/ h
sacrebleu	SacreBLEU	BLEU	Overlap		FALSE	FALSE	TRUE	39	2023.inlg-main.2	https://aclanthology.org/2023.inlg-main.2/ h
rougel	ROUGE-L	ROUGE	Overlap		FALSE	FALSE	TRUE	34	2023.inlg-main.1	https://aclanthology.org/2023.inlg-main.10/
meteor	METEOR	METEOR	Overlap		FALSE	FALSE	TRUE	27	2023.inlg-main.1	https://aclanthology.org/2023.inlg-main.1/ h
bertscore	BERTScore	BERTScore	Semantic Similarity		TRUE	FALSE	TRUE	25	2023.inlg-main.1	https://aclanthology.org/2023.inlg-main.1/ h
rouge2	ROUGE-2	ROUGE	Overlap		FALSE	FALSE	TRUE	20	2023.inlg-main.1	https://aclanthology.org/2023.inlg-main.13/
rouge1	ROUGE-1	ROUGE	Overlap		FALSE	FALSE	TRUE	19	2023.inlg-main.1	https://aclanthology.org/2023.inlg-main.13/
perplexity	Perplexity	Perplexity	Perplexity	Mimising perple	TRUE	FALSE	FALSE	19	2023.inlg-main.3	https://aclanthology.org/2023.inlg-main.31/
bleu4	BLEU-4	BLEU	Overlap		FALSE	FALSE	TRUE	14	2023.inlg-main.1	https://aclanthology.org/2023.inlg-main.1/ h
bleu1	BLEU-1	BLEU	Overlap		FALSE	FALSE	TRUE	12	2023.inlg-main.1	https://aclanthology.org/2023.inlg-main.1/ h
distinctunigrams	Distinct-1	N-gram Diversity	Text Properties	numer of distinct	FALSE	FALSE	FALSE	10	2023.inlg-main.2	https://aclanthology.org/2023.inlg-main.20/
distinctbigrams	Distinct-2	N-gram Diversity	Text Properties	number of distin	FALSE	FALSE	FALSE	10	2023.inlg-main.2	https://aclanthology.org/2023.inlg-main.20/
bleu2	BLEU-2	BLEU	Overlap		FALSE	FALSE	TRUE	9	2023.inlg-main.1	https://aclanthology.org/2023.inlg-main.1/ h
bleurt	BLEURT	BLEURT	Text Classifier		TRUE	FALSE	TRUE	9	2023.inlg-main.1	https://aclanthology.org/2023.inlg-main.12/
bartscore	BARTScore	BARTScore	Semantic Simila	Evaluates text g	TRUE	FALSE	TRUE	8	2023.inlg-main.8	https://aclanthology.org/2023.inlg-main.8/
cider	CIDEr	CIDEr	Overlap	Measures the si	FALSE	FALSE	TRUE	8	2023.inlg-main.2	https://aclanthology.org/2023.inlg-main.25/
rouge	ROUGE	ROUGE	Overlap		FALSE	FALSE	TRUE	7	2023.inlg-main.1	https://aclanthology.org/2023.inlg-main.1/ h
distincttrigrams	Distinct-3	N-gram Diversity	Text Properties	average ratio of	FALSE	FALSE	FALSE	7	2023.inlg-main.2	https://aclanthology.org/2023.inlg-main.23/
bertscoref1	BERTScore F1	BERTScore	Semantic Similarity		TRUE	FALSE	TRUE	6	P302 P1342 P2	https://aclanthology.org/2023.acl-long.674.
sari	SARI	SARI	Overlap		FALSE	TRUE	TRUE	6	2023.inlg-main.1	https://aclanthology.org/2023.inlg-main.10/
bleu3	BLEU-3	BLEU	Overlap		FALSE	FALSE	TRUE	5	2023.inlg-main.1	https://aclanthology.org/2023.inlg-main.1/ h
bertscoreprecisic	BERTScore Pre	BERTScore	Semantic Similarity		TRUE	FALSE	TRUE	5	P302 P1342 P3	https://aclanthology.org/2023.acl-long.674.
bertscorerecall	BERTScore Rec	BERTScore	Semantic Similarity		TRUE	FALSE	TRUE	5	P302 P1342 P3	https://aclanthology.org/2023.acl-long.674.

If you don’t have any details but just want to paste a link or refer to a type of metric that we are missing without specifics, there is a brainstorming tab.

Suppose there is any metric that has been particularly useful to you, or you know it is commonly used in the literature. In that case, you can add it directly to the “curated list” tab:

Name	Source	Parent Metric	Use Source	Use References	Implemented	Assigned
BLEU		BLEU	☐	☑	☑	Michael Joseph Ryan
GLEU		BLEU	☐	☑	☑	Michael Joseph Ryan
SARI-precision		SARI	☐	☑	☑	Michael Joseph Ryan
SARI-f1		SARI	☐	☑	☑	Michael Joseph Ryan
CHRF		CHRF	☐	☑	☑	Michael Joseph Ryan
TER		TER	☐	☑	☑	Michael Joseph Ryan
BERTScore-precision		BERTScore	☐	☑	☑	Michael Joseph Ryan
BERTScore-recall		BERTScore	☐	☑	☑	Michael Joseph Ryan
BERTScore-f1		BERTScore	☐	☑	☑	Michael Joseph Ryan
FKGL		FKGL	☐	☐	☑	Michael Joseph Ryan
LLM as a Judge (Likert)		LLM Judge	☐	☐	☑	Michael Joseph Ryan
LLM as a Judge (Likert + Referen		LLM Judge	☐	☑	☑	Michael Joseph Ryan
perplexity	https://aclanthol	Perplexity	☐	☐	☐	
distinctunigrams	https://aclanthol	N-gram Diversity	☐	☐	☐	
distinctbigrams	https://aclanthol	N-gram Diversity	☐	☐	☐	
bleurt	https://aclanthol	BLEURT	☐	☑	☐	
bartscore	https://aclanthol	BARTScore	☐	☑	☐	
cider	https://aclanthol	CIDEr	☐	☑	☐	
distincttrigrams	https://aclanthol	N-gram Diversity	☐	☐	☐	
bleu3	https://aclanthol	BLEU	☐	☑	☐	
mauve	https://aclanthol	MAUVE	☑	☑	☐	
selfbleu	https://aclanthol	BLEU	☐	☐	☐	
accuracy	https://aclanthol	Accuracy	☐	☑	☐	
factcc	https://aclanthol	Factuality	☐	☑	☐	
moverscore	https://aclanthol	MoverScore	☐	☑	☐	

So far, I have included metrics that were included in at least 2 ACL papers in 2024, but expanding this list and looking at metrics references from Microsoft and similar industry leaders.

If you have code ready for the metric or want to contribute the code, feel free to assign yourself!

Adding Metrics to the MetricBank

These are the steps to adding a metric to the bank:

1. Assign yourself to the to-do list: Metric ToDo
2. Ensure the repo is up to date
3. Make a new class in the `metrics` folder and extend either ReferenceBasedMetric.py or ReferenceFreeMetric.py. There is also a chance to make a ReferenceBasedMulitMetric, which essentially computes multiple metrics at the same time (such as BLEU-1, BLEU-2, etc.) but for this introduction we don't need to dive into those details.

```
class YOUR_NEW_METRIC(ReferenceBasedMetric):
   def __init__(self, name="NAME", description="DESCRIPTION HERE"):
       super().__init__(name, description)
       self.metric = YOUR_NEW_METRIC()

   def calculate(self, input, output, references=None, **kwargs):
       """
       Calculate the metric
       """
       return score
```

4. Generate documentation for the new metric using the Metric Card template and prompt: https://github.com/XenonMolecule/autometrics/blob/main/autometrics/metrics/documentation/template.md

   https://github.com/XenonMolecule/autometrics/blob/main/autometrics/metrics/documentation/prompt.txt

   NOTE: For the documentation, you will need to provide supplemental materials when using the prompt for LLM assistance. You should include a pdf of the paper, reference implementations if they exist, and any relevant details about the metric.

5. Add the documentation as a docstring for your metric. See BLEU for an example: https://github.com/XenonMolecule/autometrics/blob/main/autometrics/metrics/reference_based/BLEU.py

6. Submit a PR and I'll review it! Any tests of the metric would be a great bonus, but not strictly necessary!

Adding Task Recommendations to ToDo List

This is an area where I am especially looking for creative help. I want to test this method on very real world scenarios and ensure that it works in challenging environments. I would love to crowdsource brainstorming and suggestions of tasks! Contributing is relatively straightforward, please just add the task and description here: Task ToDo

Task	Domain	Precomputed Metrics	Size	link	Notes	Assigned	Implemented
Helpsteer	Chatbot	☐	35331	https://huggingfa	Labelled for: Helpfulness: Overall helpfulness of the response to the prompt. Correctness: Inclusion of all pertinent facts without errors. Coherence: Consistency and clarity of expression. Complexity: Intellectual depth required to write response (i.e. whether the response can be written by anyone with basic language compet Verbosity: Amount of detail included in the response, relative to what is asked for in the prompt.	Michael Jo…	☑
Helpsteer 2	Chatbot	☐	20324	https://huggingfa	Same categories as the first helpsteer	Michael Jo…	☑
SimpDA	Text Simplificatic	☐	434	https://github.cor	Adequacy, Fluency, Simplicity on a scale from 1-5	Michael Jo…	☑
Primock57	Medical	☑	1425	https://github.cor	Objective measures like time spent editing and number of errors (labelled by humans)	Michael Jo…	☑
DialSummEval	Summarization	☑	1400	https://github.cor	Coherency, Consistency, Fluency, Relevance	Michael Jo…	☑
		☐					☐
		☐					☐
		☐					☐
		☐					☐
		☐					☐
		☐					☐
		☐					☐
		☐					☐
		☐					☐

Adding Tasks to MetricBank Eval

Here are the steps for adding a Task to the MetricBank:

1. Ensure the repository is up to date
2. Navigate to `autometrics/dataset/datasets/'
3. Add in a folder for your dataset or group of datasets
4. Make a new class for your dataset; it should extend Dataset.
5. Implement the dataset loading all in the __init__ method or in functions called from __init__

```
class SimpEval(Dataset):
def __init__(self, path='./autometrics/dataset/datasets/TEST/YOUR_DATASET.csv'):
        df = pd.read_csv(path)

        # Split the reference columns into separate columns
        references = df['references'].apply(eval)

        # identify the longest reference list
        max_len = max(references.apply(len))

        for i in range(max_len):
            df[f'ref{i+1}'] = references.apply(lambda x: x[i] if len(x) > i else None)

        df.drop(columns=['references'], inplace=True)

        target_columns = ['score']
        ignore_columns = ["id","original","simple","system"]
        ignore_columns.extend([f'ref{i+1}' for i in range(max_len)])
        metric_columns = [col for col in df.columns if col not in target_columns and
col not in ignore_columns]

        name = "YOUR_DATASET"

        data_id_column = "id"
        model_id_column = "system"
        input_column = "original"
        output_column = "simple"
        reference_columns = [f'ref{i+1}' for i in range(max_len)]

        metrics = [DummyMetric(col) for col in metric_columns]
```

```
        super().__init__(df, target_columns, ignore_columns, metric_columns, name,
data_id_column, model_id_column, input_column, output_column, reference_columns,
metrics)
```

You must specify the id column, which stores the model's name, the input column, and the reference/target columns. I typically format the other dataset into a data frame and then process it. Check the other tasks for a more complete example.

6. Check that all loads properly

```
from autometrics.dataset.datasets import YOUR_DATASET
from autometrics.util.analysis import display_top_5_metrics_by_validation,
get_top_metric_by_validation = YOUR_DATASET()
```

7. Submit a PR and I'll happily review it! In the mean time feel free to test out the method on your data! I've been meaning to clean up some of the notebooks but there are plenty examples there of how to use MetricBank!

Writing Feedback

This section is still a WIP, I will pick up more on writing this after mid-February! Until then, here is an overleaf link which WILL be useful in a few weeks:
https://www.overleaf.com/project/673cf8d87789b20c36ac81bc

Brainstorming

If random ideas come to you that don't fit in the other docs, I have a brainstorming doc together just for an assortment of ideas. For instance right now I'm thinking a lot about what the metric report card outputs should look like. I would love more ideas and feedback!

Brainstorming and Notes