Research Data Management in TIRA for Reproducible Shared Tasks

March 12, Jena

Maik Fröbe, Jan Heinrich Reimer, Sean MacAvaney, Niklas Deckers, Simon Reich, Janek Bevendorff, Benno Stein, Matthias Hagen, and Martin Potthast

University of Jena University of Glasgow University of Leipzig University of Weimar

@webis_de www.webis.de

Data Management in TIRA for Reproducible Shared Tasks

What is a shared task?

Data Management in TIRA for Reproducible Shared Tasks

What is a shared task?

- Different research teams work independently on the same problem
- Fixed start and end date
- Usually, teams submit run files with the predictions of their systems

Data Management in TIRA for Reproducible Shared Tasks

What is a shared task?

- ❑ Different research teams work independently on the same problem
- ❑ Fixed start and end date
- ❑ Usually, teams submit run files with the predictions of their systems

Shared tasks shape the research in IR and NLP

- ❑ Resulting resources and annotations re-used for years

Data Management in TIRA for Reproducible Shared Tasks

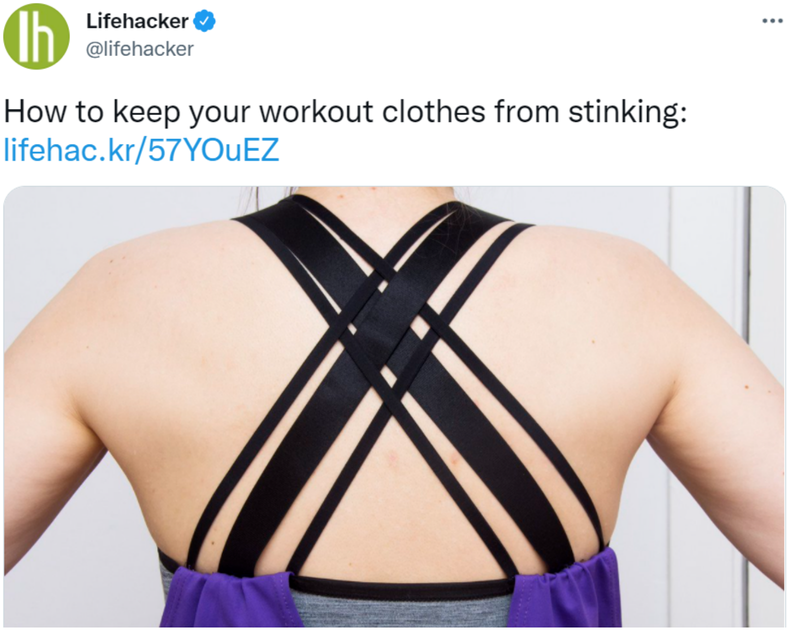
What is a shared task?

- Different research teams work independently on the same problem
- Fixed start and end date
- Usually, teams submit run files with the predictions of their systems

Shared tasks shape the research in IR and NLP

- Resulting resources and annotations re-used for years

Example Shared Task: Clickbait Spoiling (30 teams from 24 countries submitted)



Lifehacker
@lifehacker

How to keep your workout clothes from stinking:
lifehac.kr/57YOuEZ

Data Management in TIRA for Reproducible Shared Tasks

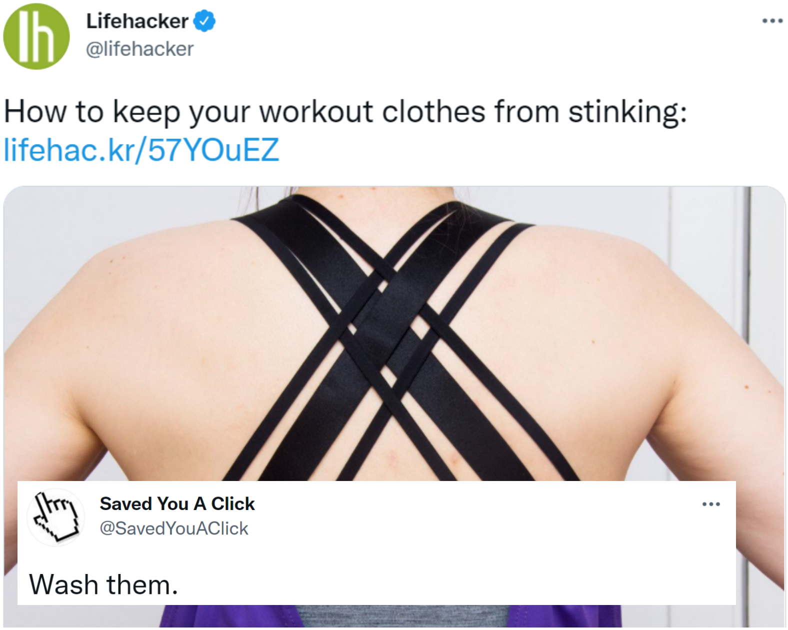
What is a shared task?

- Different research teams work independently on the same problem
- Fixed start and end date
- Usually, teams submit run files with the predictions of their systems

Shared tasks shape the research in IR and NLP

- Resulting resources and annotations re-used for years

Example Shared Task: Clickbait Spoiling (30 teams from 24 countries submitted)



Lifehacker
@lifehacker

How to keep your workout clothes from stinking:
lifehac.kr/57YOuEZ

Saved You A Click
@SavedYouAClick

Wash them.

Data Management in TIRA for Reproducible Shared Tasks

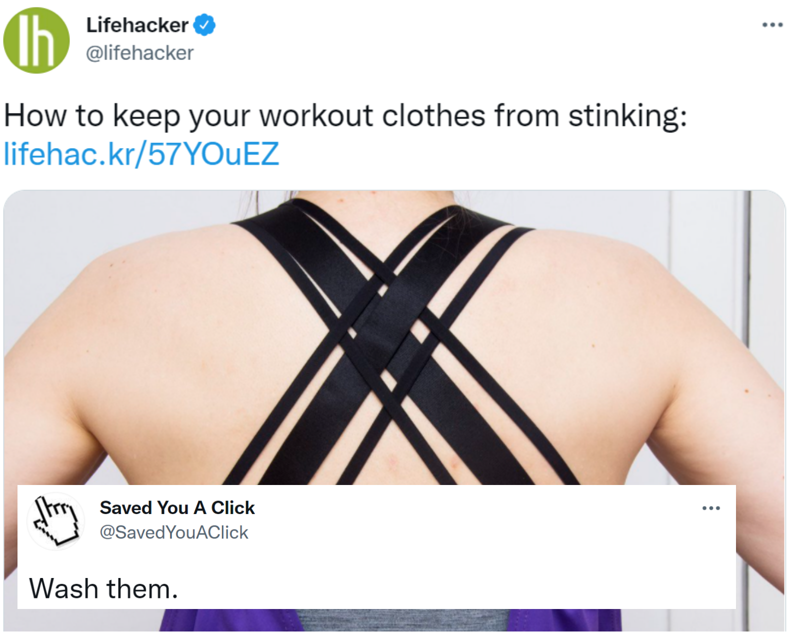
What is a shared task?

- Different research teams work independently on the same problem
- Fixed start and end date
- Usually, teams submit run files with the predictions of their systems

Shared tasks shape the research in IR and NLP

- Resulting resources and annotations re-used for years

Example Shared Task: Clickbait Spoiling (30 teams from 24 countries submitted)

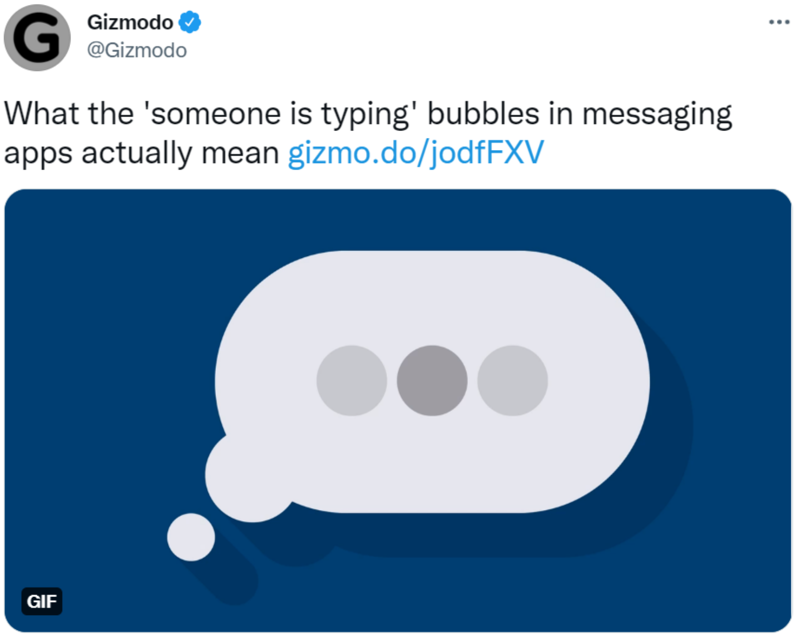


Lifehacker
@lifehacker

How to keep your workout clothes from stinking: lifehac.kr/57YOuEZ

Saved You A Click
@SavedYouAClick

Wash them.



Gizmodo
@Gizmodo

What the 'someone is typing' bubbles in messaging apps actually mean gizmo.do/jodfFXV

Data Management in TIRA for Reproducible Shared Tasks

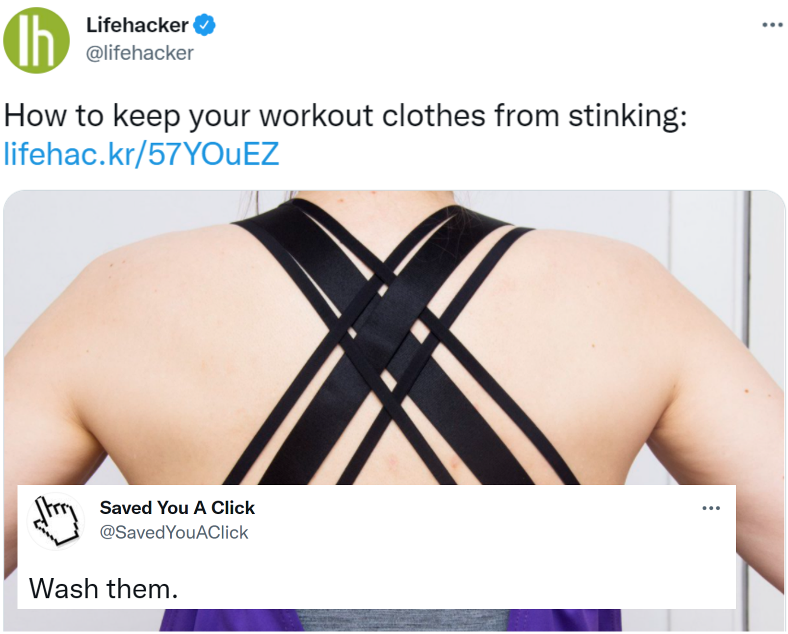
What is a shared task?

- ❑ Different research teams work independently on the same problem
- ❑ Fixed start and end date
- ❑ Usually, teams submit run files with the predictions of their systems

Shared tasks shape the research in IR and NLP

- ❑ Resulting resources and annotations re-used for years

Example Shared Task: Clickbait Spoiling (30 teams from 24 countries submitted)

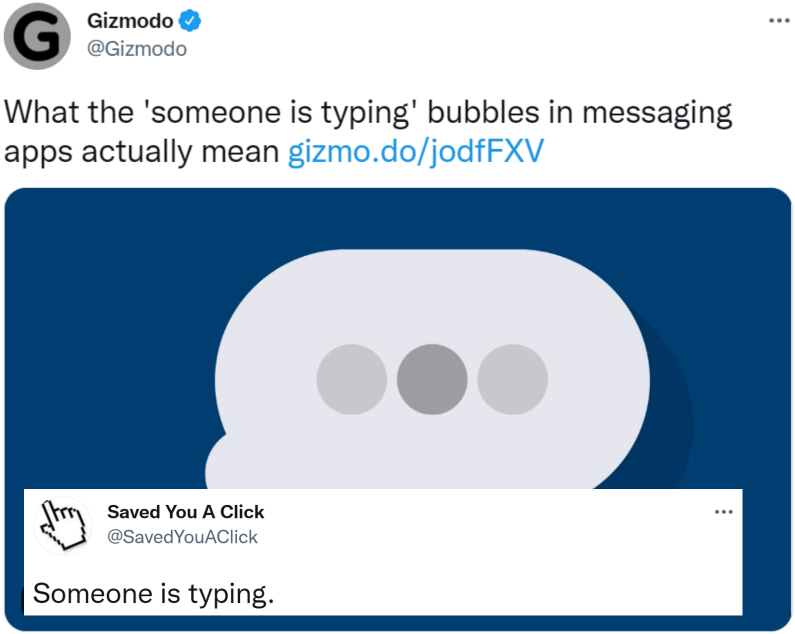


Lifehacker
@lifehacker

How to keep your workout clothes from stinking: lifehac.kr/57YOuEZ

Saved You A Click
@SavedYouAClick

Wash them.



Gizmodo
@Gizmodo

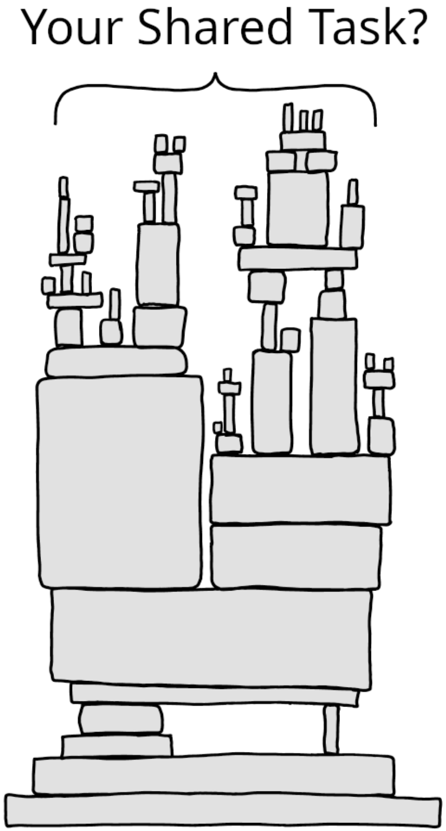
What the 'someone is typing' bubbles in messaging apps actually mean gizmo.do/jodfFXV

Saved You A Click
@SavedYouAClick

Someone is typing.

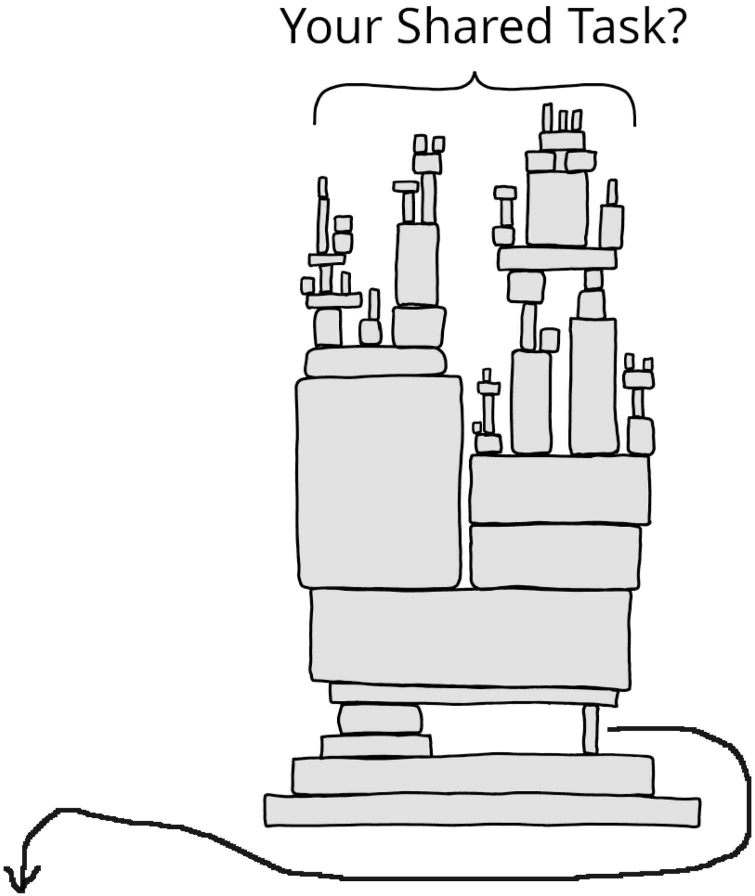
Data Management in TIRA for Reproducible Shared Tasks

Motivation



Data Management in TIRA for Reproducible Shared Tasks

Motivation

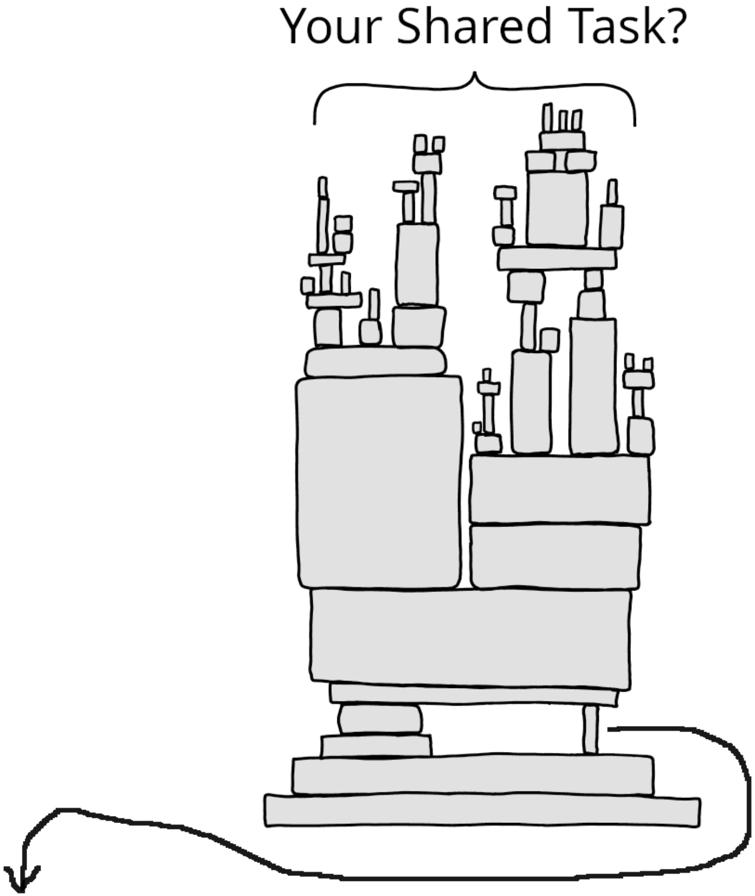


Potential problems (run submissions):
[Fuhr'21]

- ❑ Problem 1: Internal validity
- ❑ Problem 2: External validity

Data Management in TIRA for Reproducible Shared Tasks

Motivation



Potential problems (run submissions):
[Fuhr'21]

- ❑ Problem 1: Internal validity
- ❑ Problem 2: External validity
- ❑ Problem 3: Blinded experimentation with LLMs

Data Management in TIRA for Reproducible Shared Tasks

Problem 1: Internal Validity [Fuhr'21]

Goal

The hypothesis is supported by the data.

Data Management in TIRA for Reproducible Shared Tasks

Problem 1: Internal Validity [Fuhr'21]

Goal

The hypothesis is supported by the data.

Possible problems

- ❑ Wrong baseline
  [Armstrong'09,Lin'18]
- ❑ Formulate hypothesis after experiments
  [Fuhr'21]

Data Management in TIRA for Reproducible Shared Tasks

Problem 1: Internal Validity [Fuhr'21]

Goal

The hypothesis is supported by the data.

Possible problems

- Wrong baseline
  [Armstrong'09,Lin'18]
- Formulate hypothesis after experiments
  [Fuhr'21]

Possible solutions

- Centralized leaderboards
  - E.g., Run uploads to EvaluateIR
    [Armstrong'09]
- Task-specific leaderboards
  - E.g., MS MARCO, MIRACL
    [Lin'22,Zhang'22]

Data Management in TIRA for Reproducible Shared Tasks

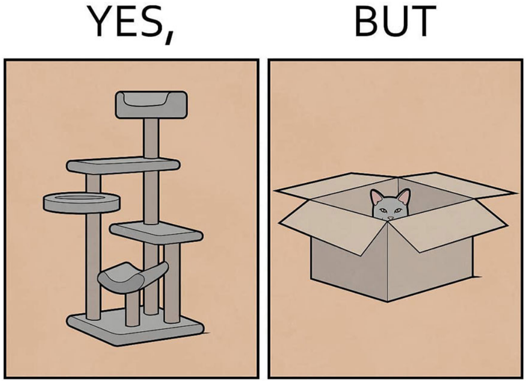
Problem 1: Internal Validity [Fuhr'21]

Goal

The hypothesis is supported by the data.

Possible problems

- Wrong baseline
  [Armstrong'09,Lin'18]
- Formulate hypothesis after experiments
  [Fuhr'21]



Possible solutions

- Centralized leaderboards
  - E.g., Run uploads to EvaluateIR
    [Armstrong'09]
- Task-specific leaderboards
  - E.g., MS MARCO, MIRACL
    [Lin'22,Zhang'22]

Data Management in TIRA for Reproducible Shared Tasks

Problem 1: Internal Validity [Fuhr’21]

Goal

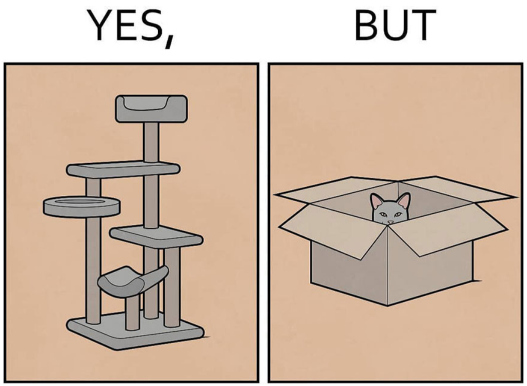
The hypothesis is supported by the data.

Possible problems

- ❑ Wrong baseline
  [Armstrong’09,Lin’18]
- ❑ Formulate hypothesis after experiments
  [Fuhr’21]

Possible solutions

- ❑ Centralized leaderboards
  - E.g., Run uploads to EvaluateIR
    [Armstrong’09]
- ❑ Task-specific leaderboards
  - E.g., MS MARCO, MIRACL
    [Lin’22,Zhang’22]



“EvaluateIR never gained traction, and a number of similar efforts following it have also floundered”
[Lin’18]

Data Management in TIRA for Reproducible Shared Tasks

Problem 2: External Validity [Fuhr'21]

Goal

Repeating an experiment on similar data yields similar observations.

Data Management in TIRA for Reproducible Shared Tasks

Problem 2: External Validity [Fuhr'21]

Goal

Repeating an experiment on similar data yields similar observations.

Possible problems

- ❑ Non-reproducible results

Data Management in TIRA for Reproducible Shared Tasks

Problem 2: External Validity [Fuhr'21]

Goal

Repeating an experiment on similar data yields similar observations.

Possible problems

- Non-reproducible results

Possible Solutions

- TREC Open Runs
  [Voorhees'16]
- Reproducibility initiatives
  - OSIRRC: Archive artifacts
    [Arguello'15,Clancy'19]
  - CENTRE: Reimplementation
    [Ferro'19,Sakai'19]
- Platforms + documentation
  - CodaLab, EvalAI, PRIMAD, STELLA, TIRA
- Meta evaluations: BEIR
  [Thakur'21]

Data Management in TIRA for Reproducible Shared Tasks

Problem 2: External Validity [Fuhr’21]

Goal

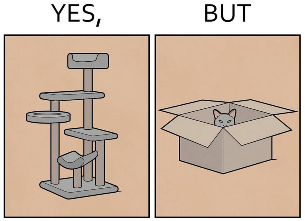
Repeating an experiment on similar data yields similar observations.

Possible problems

- Non-reproducible results

Possible Solutions

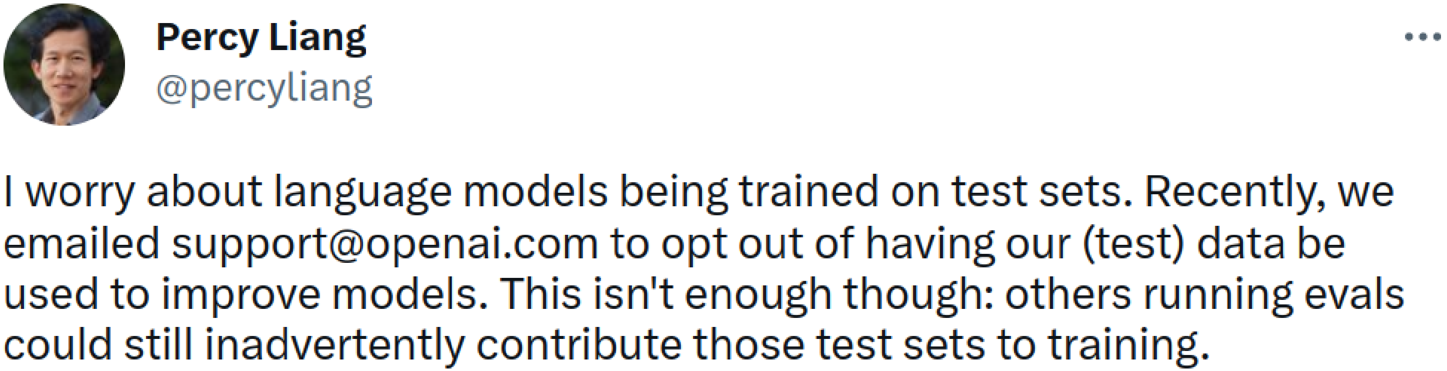
- TREC Open Runs
  [Voorhees’16]
- Reproducibility initiatives
  - OSIRRC: Archive artifacts
    [Arguello’15,Clancy’19]
  - CENTRE: Reimplementation
    [Ferro’19,Sakai’19]
- Platforms + documentation
  - CodaLab, EvalAI, PRIMAD, STELLA, TIRA
- Meta evaluations: BEIR
  [Thakur’21]



- 19 of 69 runs (Problems: 11)
- 2015: 8 systems archived
  2019: 1 system fully reproducible
  [Lin’19]
- Limited adoption of jig + CIFF
  [Clancy’19]
- Additional effort
- Evaluations on subsets
- Often sparse judgments

Data Management in TIRA for Reproducible Shared Tasks

Problem 3: Blinded Experimentation with LLMs

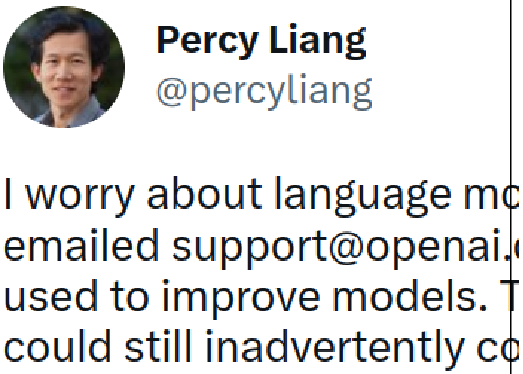


Percy Liang
@percyliang

I worry about language models being trained on test sets. Recently, we emailed support@openai.com to opt out of having our (test) data be used to improve models. This isn't enough though: others running evals could still inadvertently contribute those test sets to training.

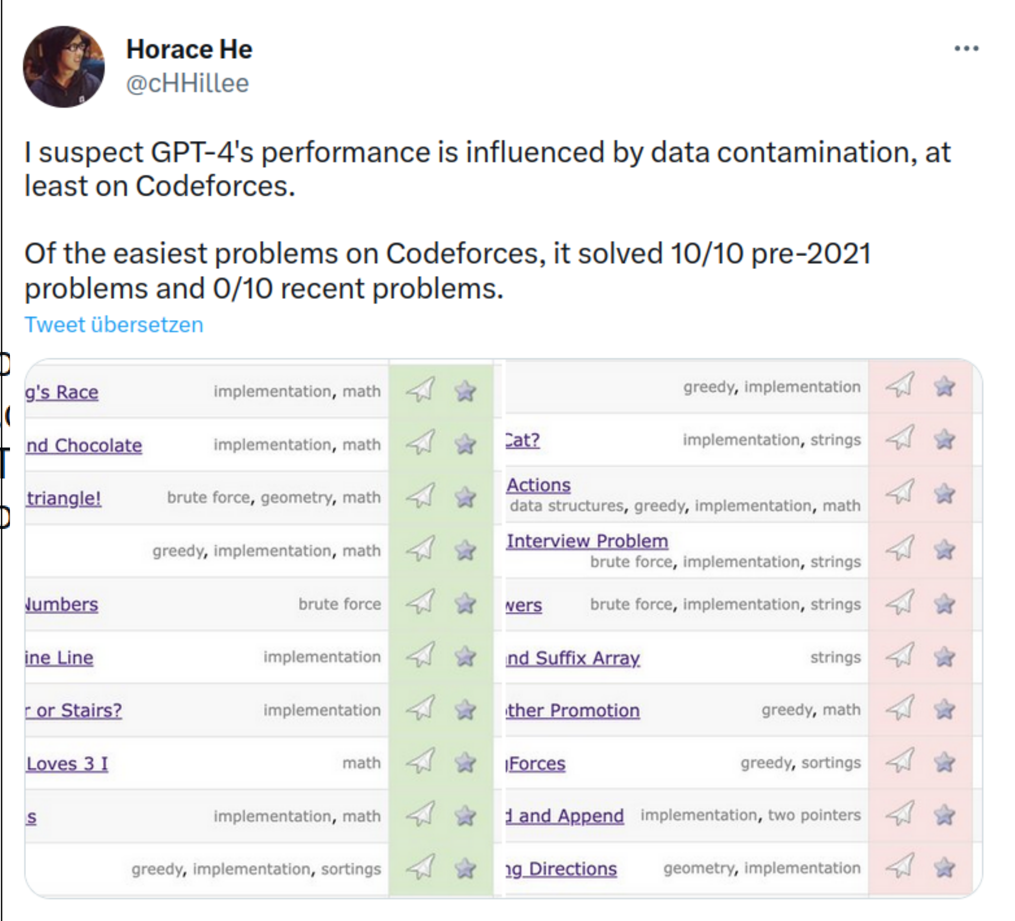
Data Management in TIRA for Reproducible Shared Tasks

Problem 3: Blinded Experimentation with LLMs



Percy Liang
@percyliang

I worry about language mo
emailed support@openai.
used to improve models. T
could still inadvertently co



Horace He
@cHHillee

I suspect GPT-4's performance is influenced by data contamination, at least on Codeforces.

Of the easiest problems on Codeforces, it solved 10/10 pre-2021 problems and 0/10 recent problems.

g's Race	implementation, math		greedy, implementation
nd Chocolate	implementation, math	Cat?	implementation, strings
triangle!	brute force, geometry, math	Actions	data structures, greedy, implementation, math
	greedy, implementation, math	Interview Problem	brute force, implementation, strings
Numbers	brute force	wers	brute force, implementation, strings
ine Line	implementation	nd Suffix Array	strings
r or Stairs?	implementation	ther Promotion	greedy, math
Loves 3 I	math	Forces	greedy, sortings
s	implementation, math	d and Append	implementation, two pointers
	greedy, implementation, sortings	ng Directions	geometry, implementation

TIRA to the Rescue?

Reproducible Shared Tasks with TIRA

Evolution of TIRA

[Gollub'12,Potthast'19,Fröbe'23]

- 2005–2011: Pipelines, eval. run submissions, manual software submissions
- 2012–2022: Software submissions with virtual machines
- 2023–today: Immutable software submissions with Docker + Git CI/CD
  - Shared task = git repository
  - Software execution = commit

Reproducible Shared Tasks with TIRA

Evolution of TIRA

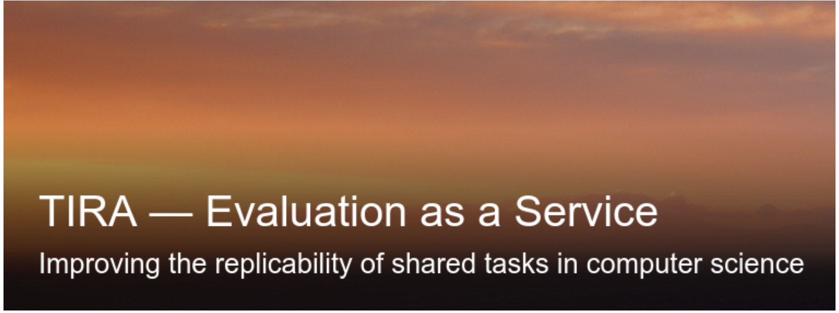
[Gollub'12,Potthast'19,Fröbe'23]

- 2005–2011: Pipelines, eval. run submissions, manual software submissions
- 2012–2022: Software submissions with virtual machines
- 2023–today: Immutable software submissions with Docker + Git CI/CD
  - Shared task = git repository
  - Software execution = commit

Procedure:

1. Implement approach in Docker image
2. Upload image to dedicated image registry in TIRA
3. Your approach is executed in a Kubernetes cluster via a commit

`http://tira.io`



TIRA — Evaluation as a Service

Improving the replicability of shared tasks in computer science

Benefits of TIRA

Blinded Experimentation

- ❑ Software executed in sandbox: No internet connection
- ❑ 2 types of datasets:

Type	Blinded	Unblinding	Feedback
Validation	Nothing	Direct	Everything
Test	Everything	Manual	✓vs ✗

Benefits of TIRA

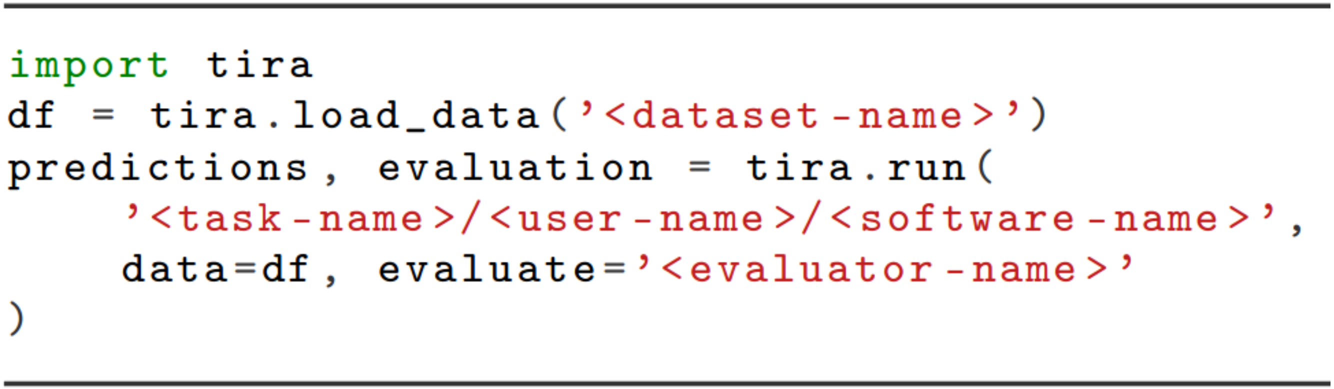
Blinded Experimentation

- Software executed in sandbox: No internet connection
- 2 types of datasets:

Type	Blinded	Unblinding	Feedback
Validation	Nothing	Direct	Everything
Test	Everything	Manual	✓vs ✗

Repeat, Replicate, and Reproduce in One Line of Code

- Git repository of the shared task can be published after the task

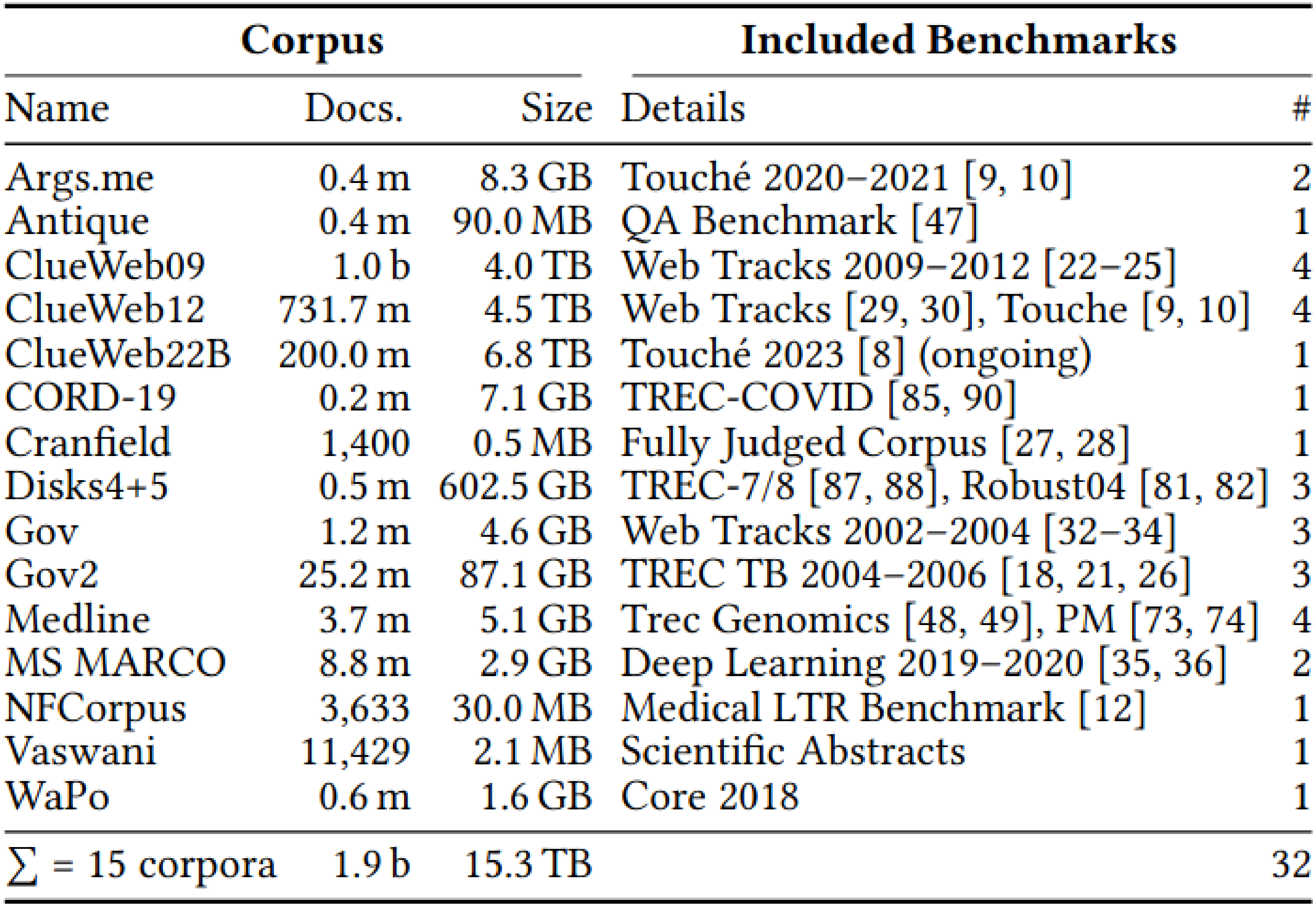


```
import tira
df = tira.load_data(’<dataset-name>’)
predictions, evaluation = tira.run(
    ’<task-name>/<user-name>/<software-name>’,
    data=df, evaluate=’<evaluator-name>’
)
```

Research Data Management in TIRA

Interoperability to Improve Internal and External Validity (1)

- Standardized access and integration of 32 IR test collections to TIRA
- Models can be transferred to new corpora ⇒ improves external validity

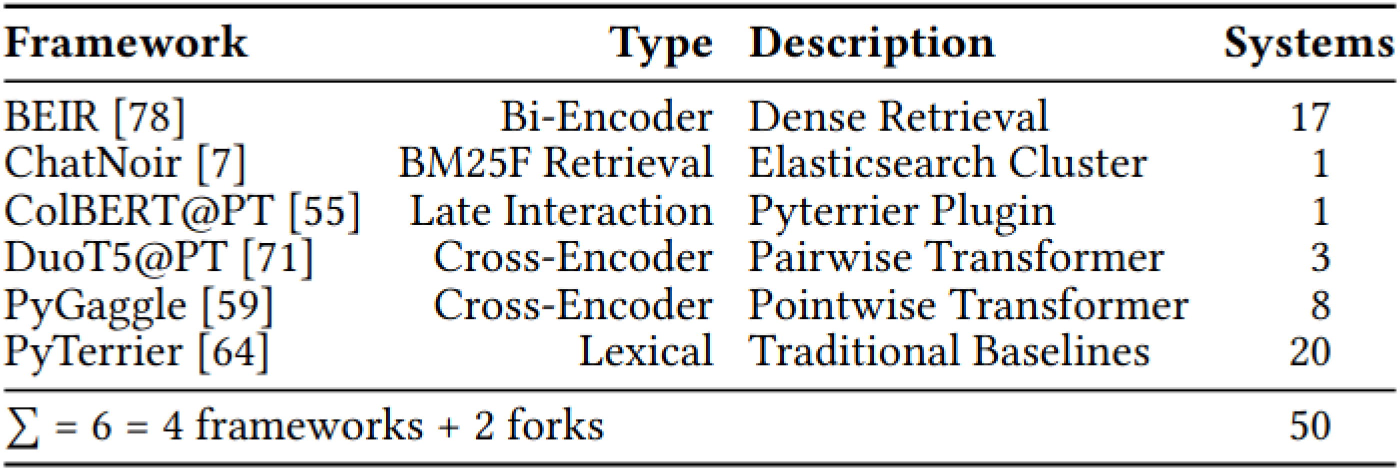


Corpus			Included Benchmarks	
Name	Docs.	Size	Details	#
Args.me	0.4 m	8.3 GB	Touché 2020–2021 [9, 10]	2
Antique	0.4 m	90.0 MB	QA Benchmark [47]	1
ClueWeb09	1.0 b	4.0 TB	Web Tracks 2009–2012 [22–25]	4
ClueWeb12	731.7 m	4.5 TB	Web Tracks [29, 30], Touche [9, 10]	4
ClueWeb22B	200.0 m	6.8 TB	Touché 2023 [8] (ongoing)	1
CORD-19	0.2 m	7.1 GB	TREC-COVID [85, 90]	1
Cranfield	1,400	0.5 MB	Fully Judged Corpus [27, 28]	1
Disks4+5	0.5 m	602.5 GB	TREC-7/8 [87, 88], Robust04 [81, 82]	3
Gov	1.2 m	4.6 GB	Web Tracks 2002–2004 [32–34]	3
Gov2	25.2 m	87.1 GB	TREC TB 2004–2006 [18, 21, 26]	3
Medline	3.7 m	5.1 GB	Trec Genomics [48, 49], PM [73, 74]	4
MS MARCO	8.8 m	2.9 GB	Deep Learning 2019–2020 [35, 36]	2
NFCorpus	3,633	30.0 MB	Medical LTR Benchmark [12]	1
Vaswani	11,429	2.1 MB	Scientific Abstracts	1
WaPo	0.6 m	1.6 GB	Core 2018	1
$\sum$ = 15 corpora	1.9 b	15.3 TB		32

Research Data Management in TIRA

Interoperability to Improve Internal and External Validity (2)

- 50 Transferrable Retrieval Models in TIRA
- Selecting suitable baseline → improves internal validity



Framework	Type	Description	Systems
BEIR [78]	Bi-Encoder	Dense Retrieval	17
ChatNoir [7]	BM25F Retrieval	Elasticsearch Cluster	1
ColBERT@PT [55]	Late Interaction	Pyterrier Plugin	1
DuoT5@PT [71]	Cross-Encoder	Pairwise Transformer	3
PyGaggle [59]	Cross-Encoder	Pointwise Transformer	8
PyTerrier [64]	Lexical	Traditional Baselines	20
$\sum$ = 6 = 4 frameworks + 2 forks			50

Research Data Management in TIRA

Goal

- Remove all dependencies to our infrastructure after the shared task
- Maintenance reduced to active shared tasks

During the Shared Task:

- We maintain and help
- Docker images in private registry
- Input data and outputs in CephFS

After the Shared Task

- Goal: Post-hoc experiments and analysis even when our cluster is down
- Docker images to Dockerhub
- Shared task repository to Github
- Input data to Zenodo
- All outputs to Zenodo + task-specific Python wrapper
  - Simplifies replicability experiments + analysis

Conclusion

TIRA allows shared tasks on confidential data with software submissions

- ❑ Improved Reproducibility
- ❑ Blinded Experimentation

Interoperability for better benefit/effort ratio

- ❑ One software submission, evaluation on many datasets
- ❑ Evaluate on datasets to which you dont have access

Conclusion

TIRA allows shared tasks on confidential data with software submissions

- ❑ Improved Reproducibility
- ❑ Blinded Experimentation

Interoperability for better benefit/effort ratio

- ❑ One software submission, evaluation on many datasets
- ❑ Evaluate on datasets to which you dont have access

Future Work

- ❑ Upcoming evaluation campaigns co-located with major IR Conferences
  - – CLEF’24, ECIR’24, SIGIR’24

Conclusion

TIRA allows shared tasks on confidential data with software submissions

- Improved Reproducibility
- Blinded Experimentation

Interoperability for better benefit/effort ratio

- One software submission, evaluation on many datasets
- Evaluate on datasets to which you dont have access

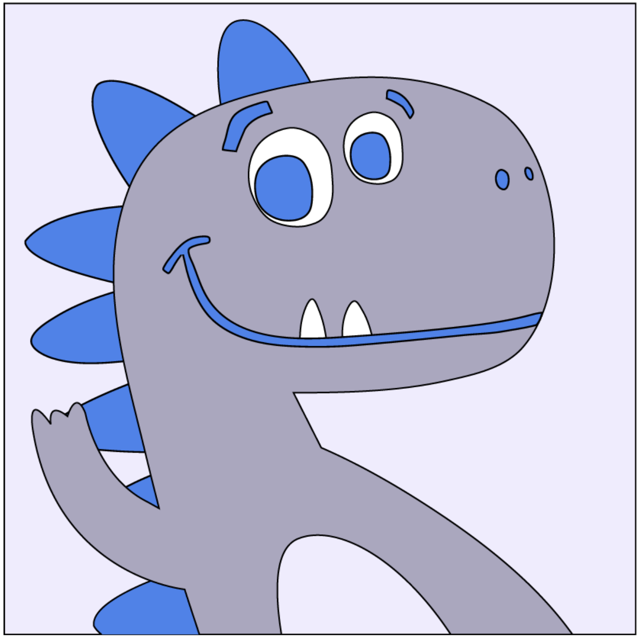
Future Work

- Upcoming evaluation campaigns co-located with major IR Conferences
  - CLEF’24, ECIR’24, SIGIR’24

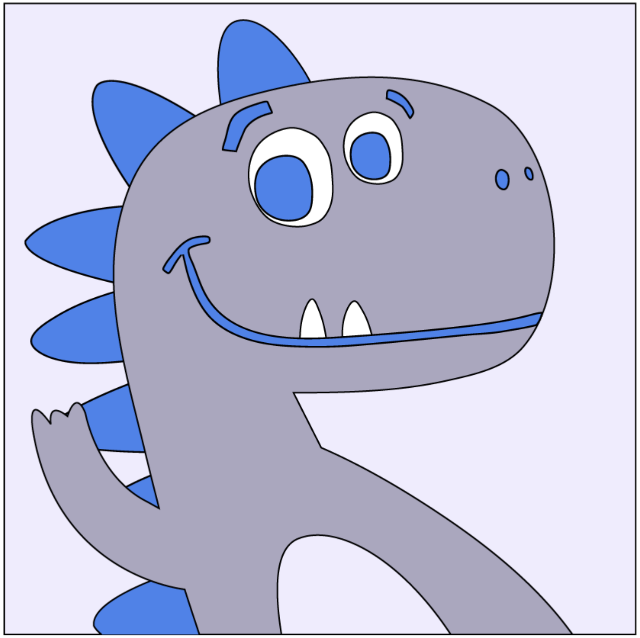
github.com/tira-io/tira

Thank You!

Example: TIREx



Example: TIREx



TIREx does “one thing”: Integrate Existing Tools

TIRA

- ❑ Reproducible shared tasks: Software submissions + blinded experiments

ir_datasets

- ❑ Unified + random data access: Documents + queries + rel. Judgments

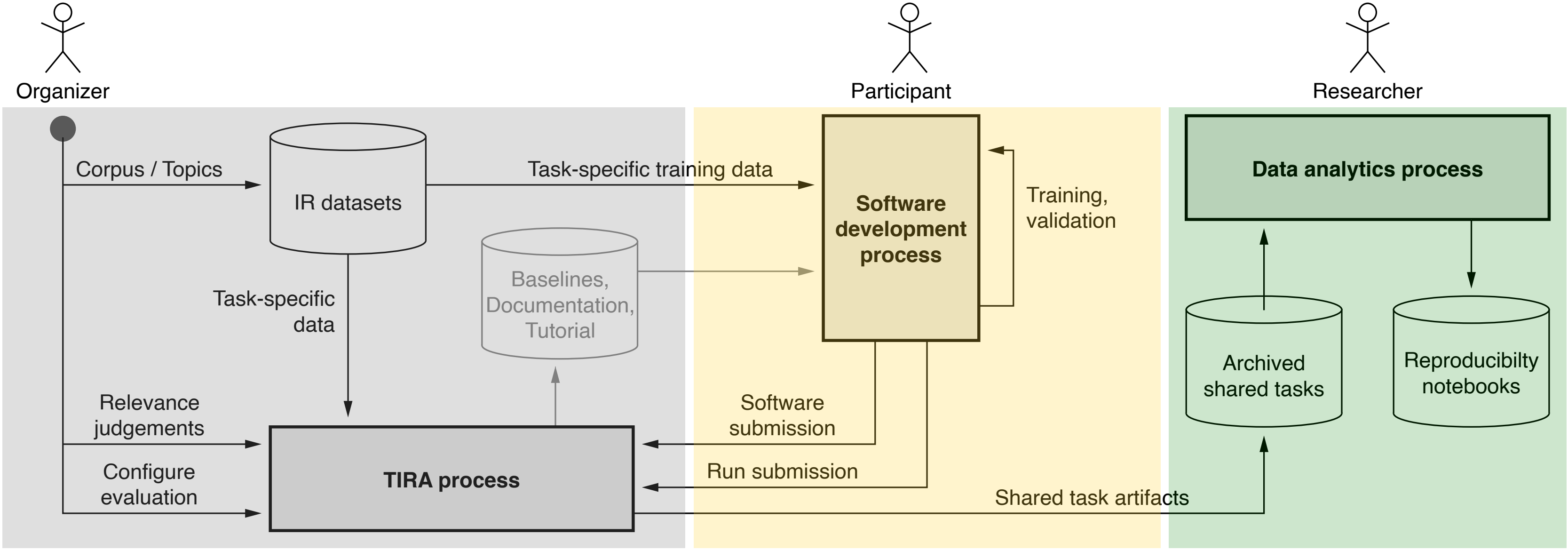
PyTerrier

- ❑ Declarative reproducibility pipelines

TIREx: Overview

- ❑ Organizer provides (private) docker image with ir_datasets integration
- ❑ Participants provide docker images with retrieval approaches

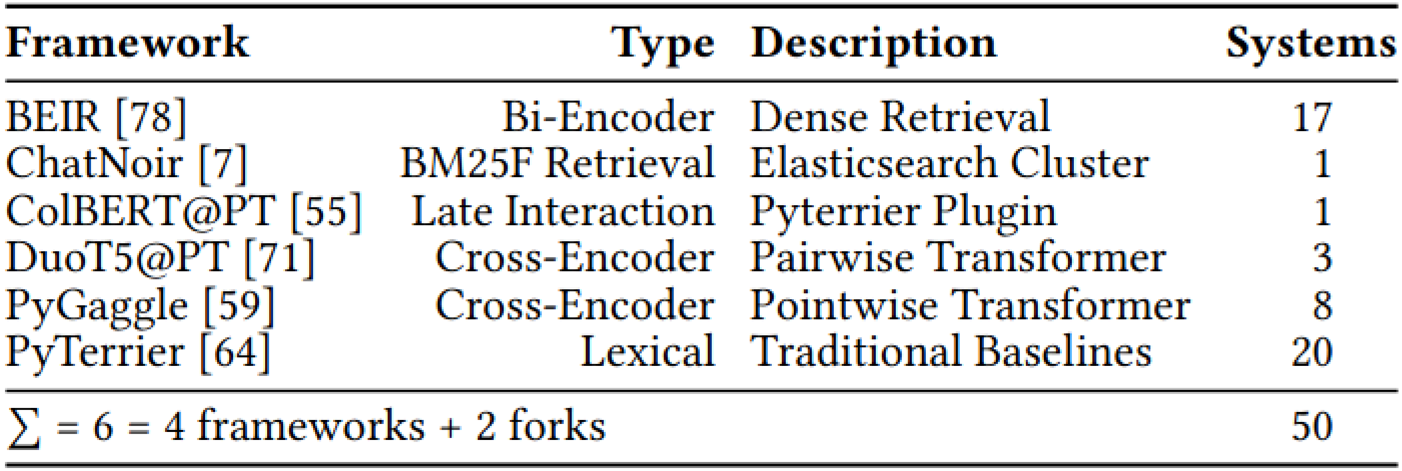
Covers a shared task end-to-end



TIREx: Feasibility Study

50 Transferrable Retrieval Models in TIRA

- Derived from tira-starters from 4 starters
- Retrieve against default text in ir_datasets
- Selecting suitable baseline → improves internal validity
- Diversification of pools for shared tasks with few participants

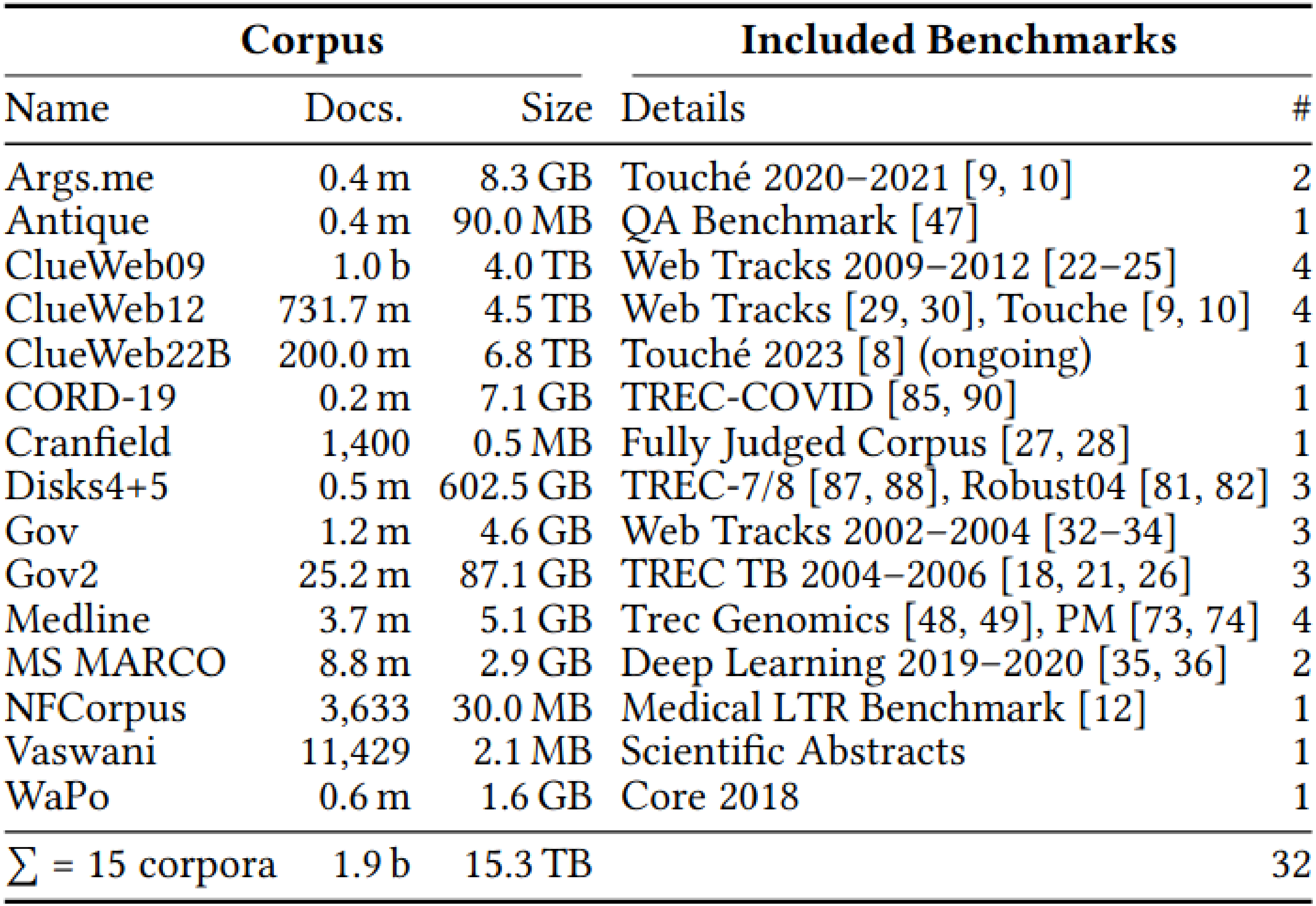


Framework	Type	Description	Systems
BEIR [78]	Bi-Encoder	Dense Retrieval	17
ChatNoir [7]	BM25F Retrieval	Elasticsearch Cluster	1
ColBERT@PT [55]	Late Interaction	Pyterrier Plugin	1
DuoT5@PT [71]	Cross-Encoder	Pairwise Transformer	3
PyGaggle [59]	Cross-Encoder	Pointwise Transformer	8
PyTerrier [64]	Lexical	Traditional Baselines	20
$\sum = 6 = 4$ frameworks + 2 forks			50

TIREx: Feasibility Study

32 Exchangeable Benchmarks in TIRA

- Models can be transferred to new corpora ⇒ improves external validity



Corpus			Included Benchmarks	
Name	Docs.	Size	Details	#
Args.me	0.4 m	8.3 GB	Touché 2020–2021 [9, 10]	2
Antique	0.4 m	90.0 MB	QA Benchmark [47]	1
ClueWeb09	1.0 b	4.0 TB	Web Tracks 2009–2012 [22–25]	4
ClueWeb12	731.7 m	4.5 TB	Web Tracks [29, 30], Touche [9, 10]	4
ClueWeb22B	200.0 m	6.8 TB	Touché 2023 [8] (ongoing)	1
CORD-19	0.2 m	7.1 GB	TREC-COVID [85, 90]	1
Cranfield	1,400	0.5 MB	Fully Judged Corpus [27, 28]	1
Disks4+5	0.5 m	602.5 GB	TREC-7/8 [87, 88], Robust04 [81, 82]	3
Gov	1.2 m	4.6 GB	Web Tracks 2002–2004 [32–34]	3
Gov2	25.2 m	87.1 GB	TREC TB 2004–2006 [18, 21, 26]	3
Medline	3.7 m	5.1 GB	Trec Genomics [48, 49], PM [73, 74]	4
MS MARCO	8.8 m	2.9 GB	Deep Learning 2019–2020 [35, 36]	2
NFCorpus	3,633	30.0 MB	Medical LTR Benchmark [12]	1
Vaswani	11,429	2.1 MB	Scientific Abstracts	1
WaPo	0.6 m	1.6 GB	Core 2018	1
$\sum$ = 15 corpora	1.9 b	15.3 TB		32

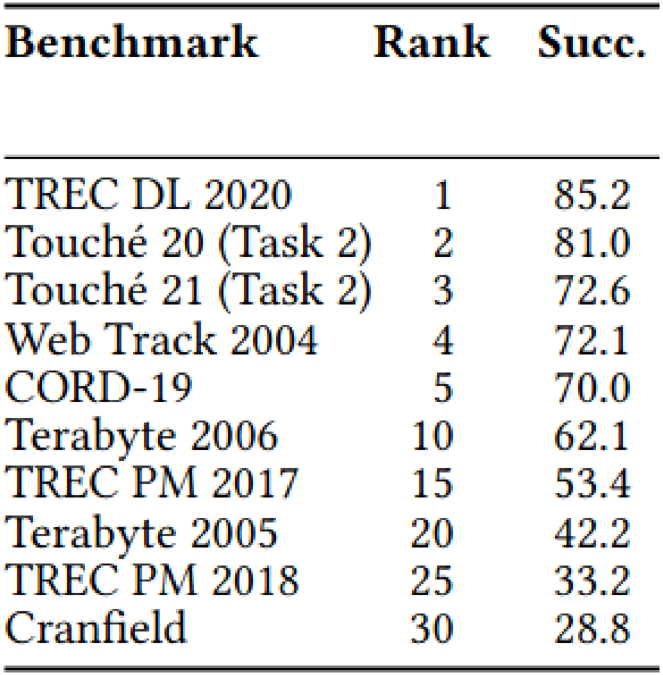
TIREx: Feasibility Study

Initial Leaderboards: 1600 runs

- Running all 50 models on all benchmarks took 1 Week
- See https://github.com/tira-io/ir-experiment-platform
- Additional use-cases: LTR, QPP, etc.

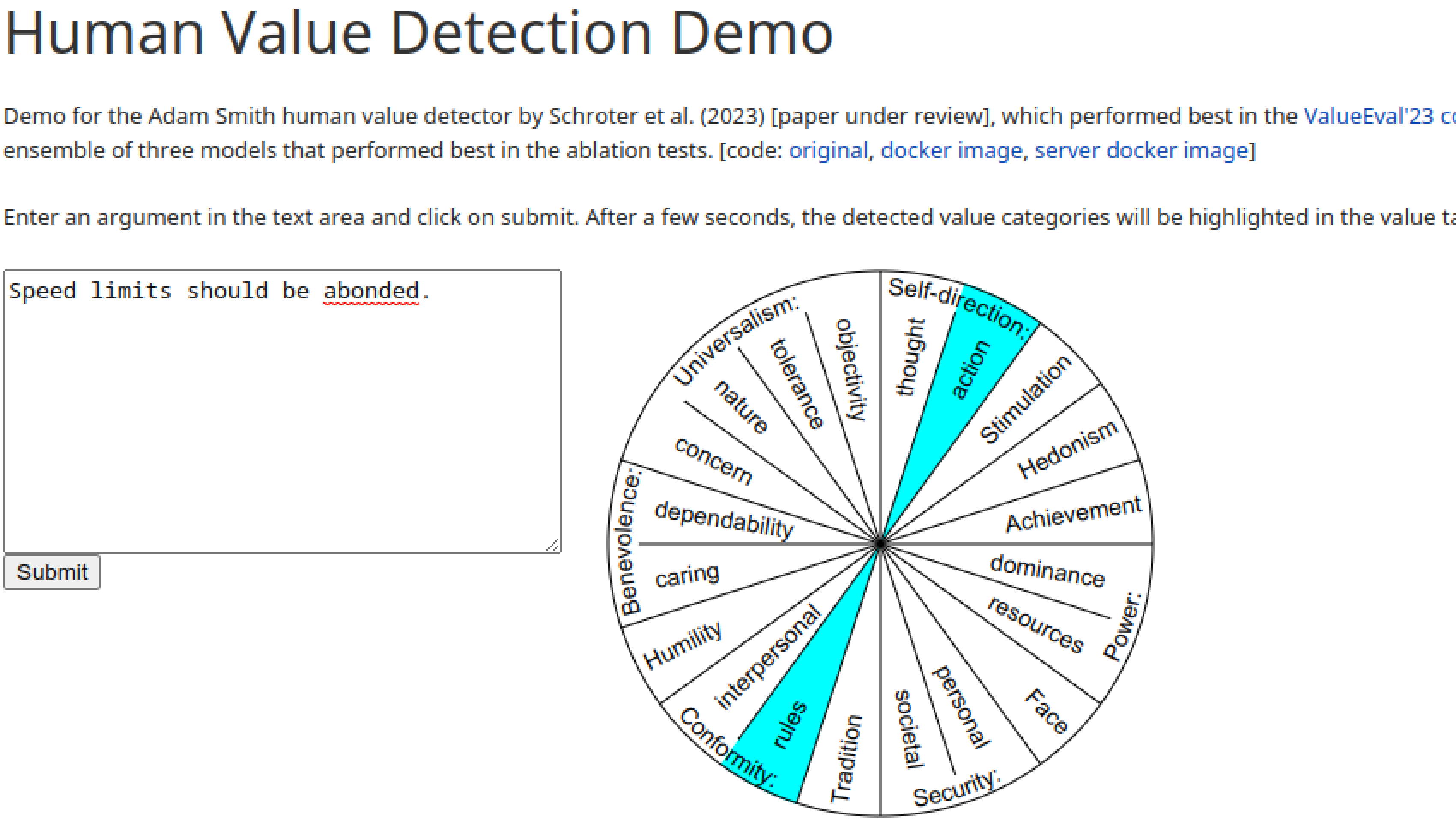
Teaser of results:

- Observe system preferences on TREC DL 2019
- Use repro_eval to measure the proportion of reproducible preferences [Breuer'20,Breuer'21]



Benchmark	Rank	Succ.
TREC DL 2020	1	85.2
Touché 20 (Task 2)	2	81.0
Touché 21 (Task 2)	3	72.6
Web Track 2004	4	72.1
CORD-19	5	70.0
Terabyte 2006	10	62.1
TREC PM 2017	15	53.4
Terabyte 2005	20	42.2
TREC PM 2018	25	33.2
Cranfield	30	28.8

Backup: SemEval’23 ValueEval Demo (1)



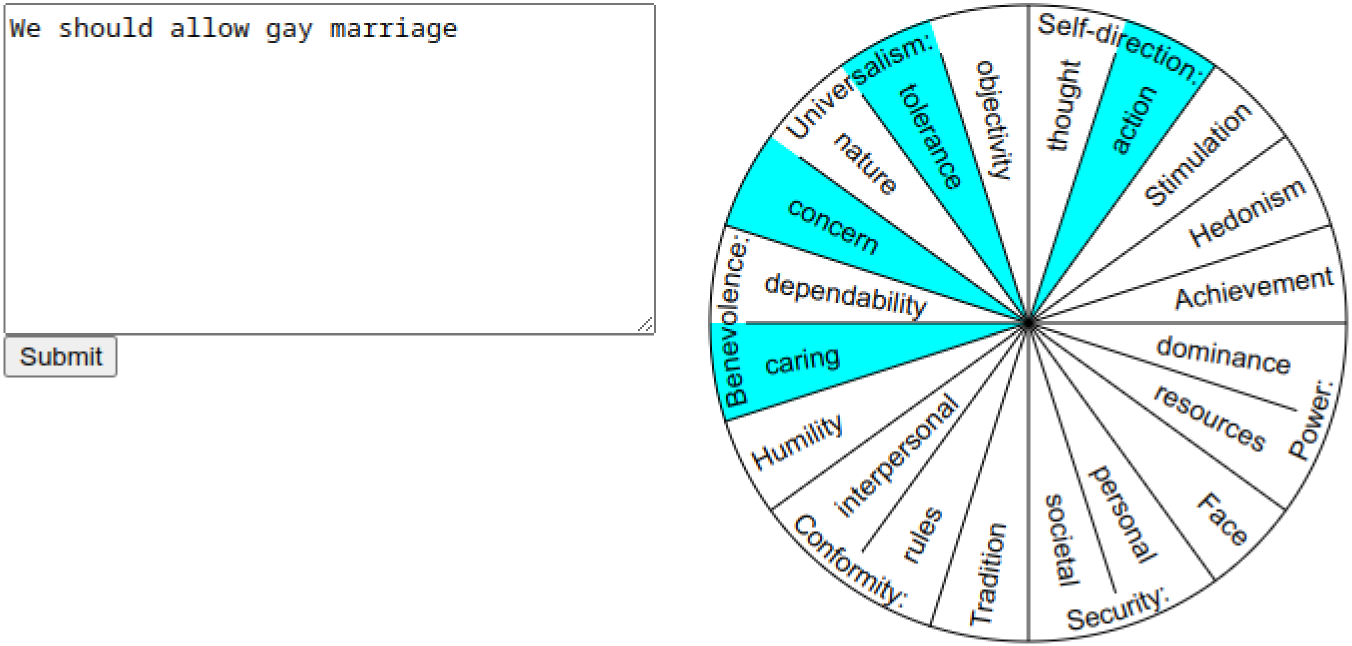
Human Value Detection Demo

Demo for the Adam Smith human value detector by Schroter et al. (2023) [paper under review], which performed best in the ValueEval'23 co
ensemble of three models that performed best in the ablation tests. [code: original, docker image, server docker image]

Enter an argument in the text area and click on submit. After a few seconds, the detected value categories will be highlighted in the value ta

```
Speed limits should be abonded.
```

Backup: SemEval’23 ValueEval Demo (2)



We should allow gay marriage

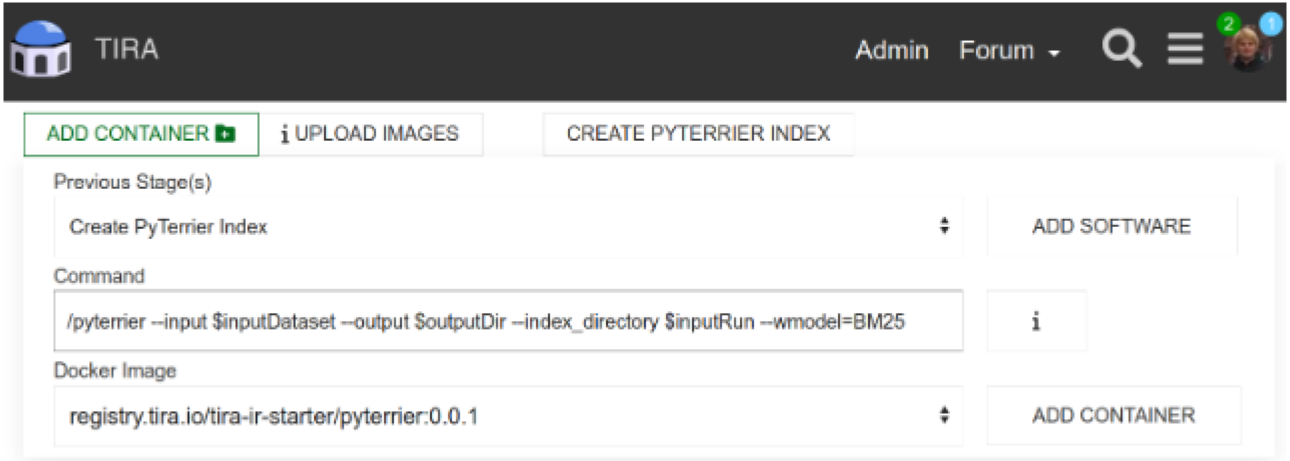
Backup: Limitations

- Computational resources.
  Potential Solution:
  - Hybrid submissions: Run upload, Software submission only for plausibility checks
  -
  - OSF infrastructure
- How to avoid big ensembles?
- Evaluation measures required that combine efficiency with effectiveness?
- New iteration of the IRF?

Backup: Use in Teaching

- Cover the “full cycle” with students in IR exercises?
  - We do this next term

Backup: Definition of Multi-Stage Software



Previous Stage(s): Create PyTerrier Index

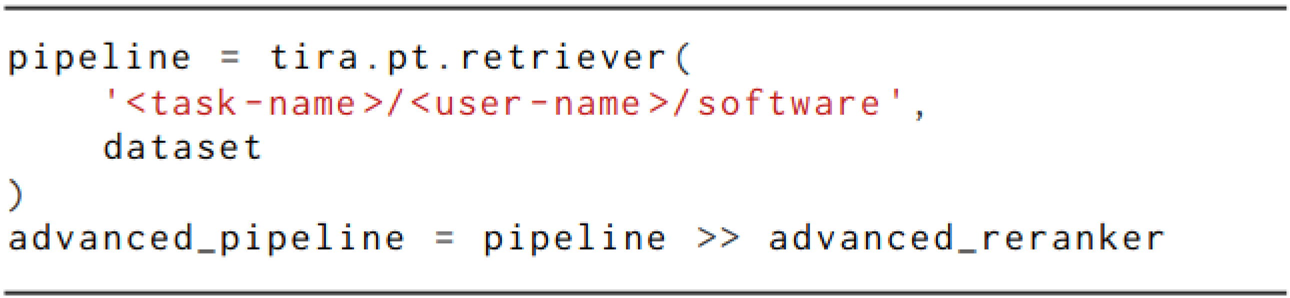
Command:

```
/pyterrier --input $inputDataset --output $outputDir --index_directory $inputRun --wmodel=BM25
```

Docker Image: registry.tira.io/tira-ir-starter/pyterrier:0.0.1

Figure 3: The definition of a full-rank retrieval software in TIRA that consists of two modularized components.

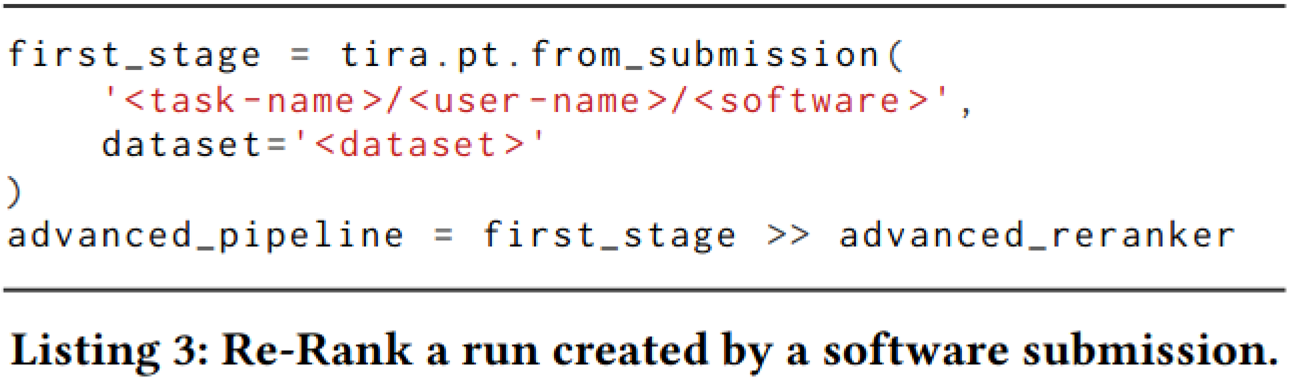
Backup: Full-Rank



```
pipeline = tira.pt.retriever(
    '<task-name>/<user-name>/software',
    dataset
)
advanced_pipeline = pipeline >> advanced_reranker
```

Listing 1: Full-Rank Retrieval from a complete corpus.

Backup: Load Submissions



```
first_stage = tira.pt.from_submission(
    '<task-name>/<user-name>/<software>',
    dataset='<dataset>'
)
advanced_pipeline = first_stage >> advanced_reranker
```

Listing 3: Re-Rank a run created by a software submission.