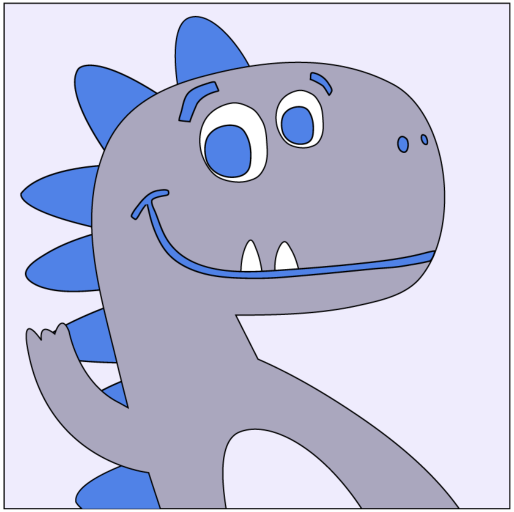
TIREx: The Information Retrieval Experiment Platform

Towards Reproducible Shared Tasks in IR



Glasgow IR Seminar, 31th March, 2023

Maik Fröbe, Jan Heinrich Reimer, Sean MacAvaney, Niklas Deckers, Simon Reich, Janek Bevendorff, Matti Wiegmann, Nikolay Kolyada, Bastian Grahm, Theresa Elstner, Frank Loebe, Tim Gollub, Benno Stein, Matthias Hagen, and Martin Potthast

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

@webis_de   www.webis.de

TIREx: The Information Retrieval Experiment Platform

Motivation

Michael Granitzer

Leiter OpenWebSearch.eu

"I want to choose my search engine like my daily newspaper"

TIREx: The Information Retrieval Experiment Platform

Motivation

Open Search Foundation

- Joint EU project
- Open Web Index to foster competition
- Shared tasks and data challenges planned

TIREx: The Information Retrieval Experiment Platform

Motivation

Michael Granitzer

Leiter OpenWebSearch.eu

"I want to choose my search engine like my daily newspaper"

Open Search Foundation

- ❏ Joint EU project
- ❏ Open Web Index to foster competition
- ❏ **Shared tasks** and data challenges planned

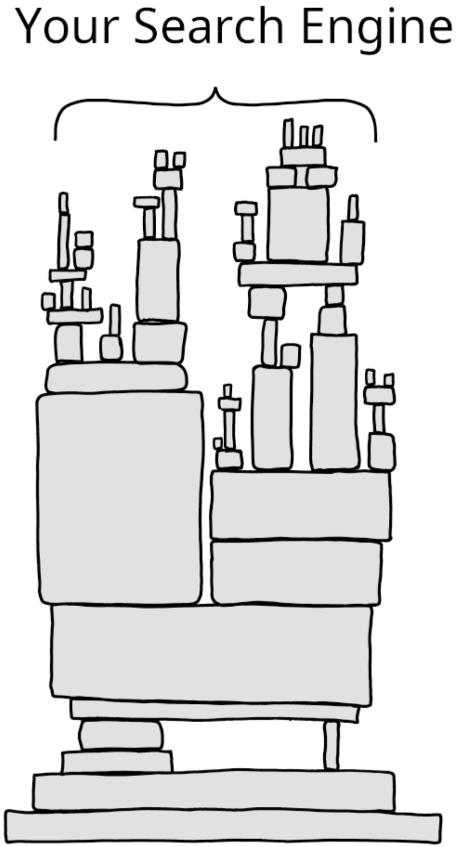
TIREx: The Information Retrieval Experiment Platform

Best Case

TIREx: The Information Retrieval Experiment Platform

Best Case

Worst Case

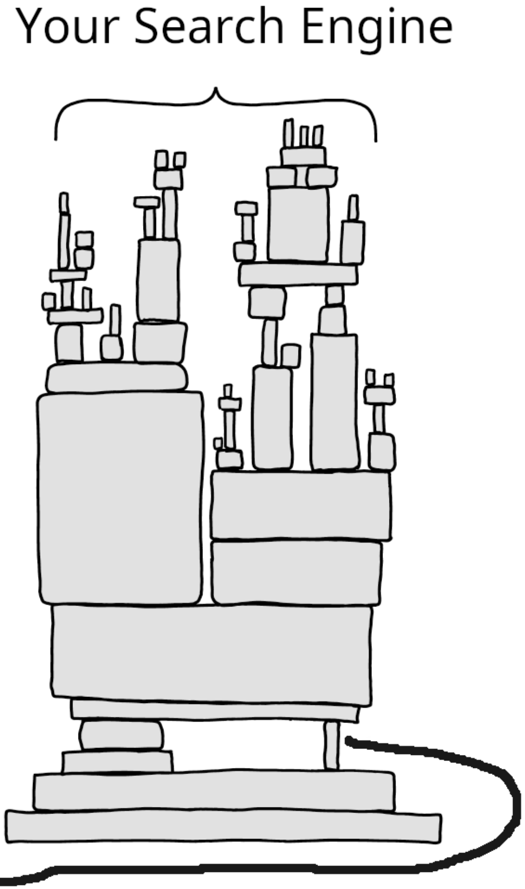


© Webis 2023

TIREx: The Information Retrieval Experiment Platform

Best Case

Worst Case

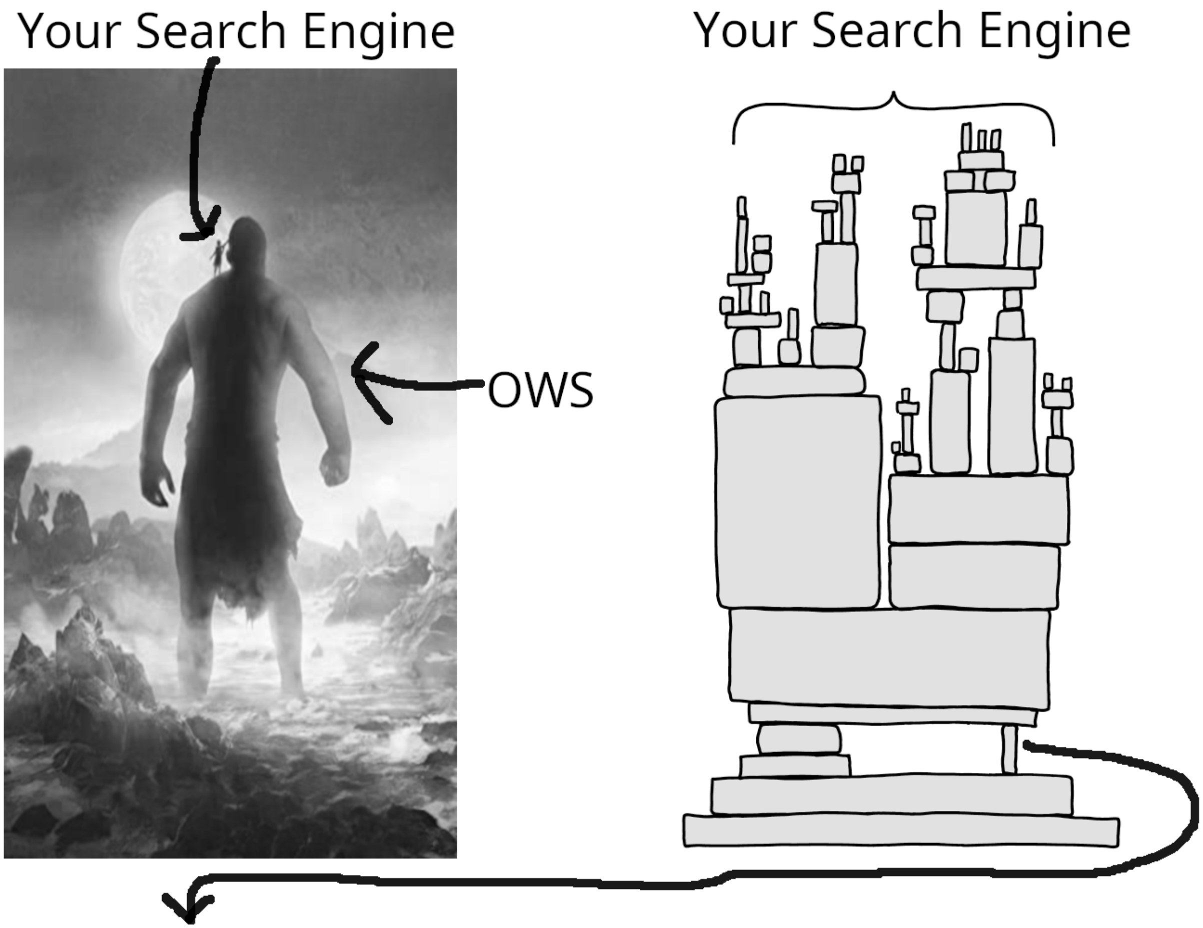


Potential problems:
[Fuhr'21]

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

TIREx: The Information Retrieval Experiment Platform

Best Case Worst Case



Potential problems:

[Fuhr'21]

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

TIREx: The Information Retrieval Experiment Platform

Problem 1: Internal Validity [Fuhr'21]

Goal

The hypothesis is supported by the data.

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]

TIREx: The Information Retrieval Experiment Platform

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]

© Webis 2023

TIREx: The Information Retrieval Experiment Platform

Problem 1: Internal Validity [Fuhr'21]

Goal

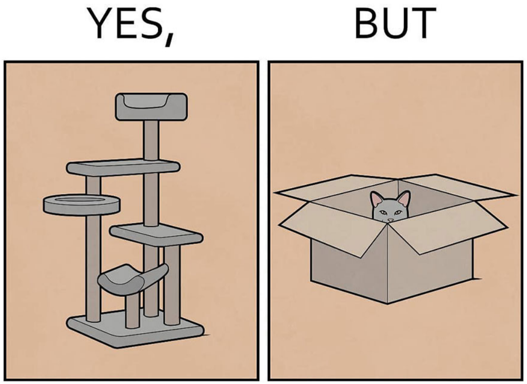
The hypothesis is supported by the data.

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]



© Webis 2023

TIREx: The Information Retrieval Experiment Platform

Problem 1: Internal Validity [Fuhr’21]

Goal

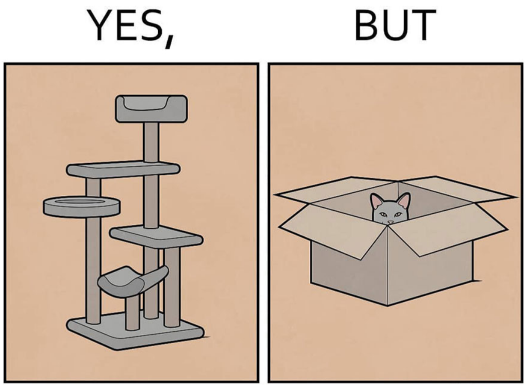
The hypothesis is supported by the data.

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]

© Webis 2023

TIREx: The Information Retrieval Experiment Platform

Problem 2: External Validity [Fuhr'21]

Goal

Repeating an experiment on similar data yields similar observations.

© Webis 2023

TIREx: The Information Retrieval Experiment Platform

Problem 2: External Validity [Fuhr'21]

Goal

Repeating an experiment on similar data yields similar observations.

Possible problems

- Non-reproducible results

© Webis 2023

TIREx: The Information Retrieval Experiment Platform

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]

©Webis 2023

TIREx: The Information Retrieval Experiment Platform

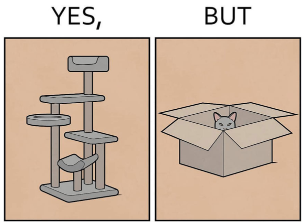
Problem 2: External Validity [Fuhr’21]

Goal

Repeating an experiment on similar data yields similar observations.

Possible problems

- Non-reproducible results



Possible Solutions

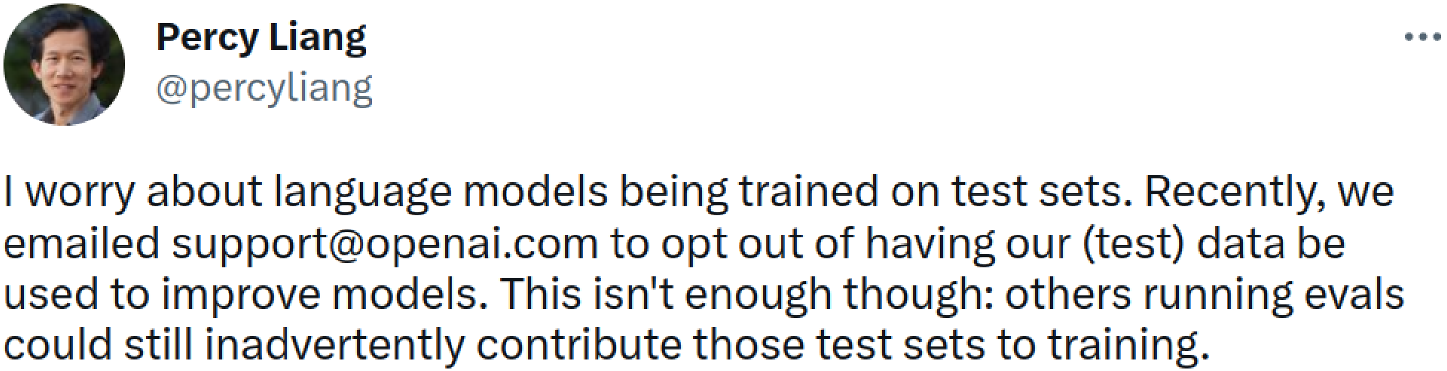
- TREC Open Runs
  [Voorhees’16]
- Reproducibility initiatives
  - OSIRRC: Archive artifacts
    [Arguello’15,Clancy’19]
  - CENTRE: Reimplementation
    [Ferro’19,Sakai’19]
- Platforms + documentation
  - CodaLab, EvalAI, PRIMAD, STELLA, TIRA
- Meta evaluations: BEIR
  [Thakur’21]

- 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

© Webis 2023

TIREx: The Information Retrieval Experiment Platform

Problem 3: Blinded Experimentation with LLMs

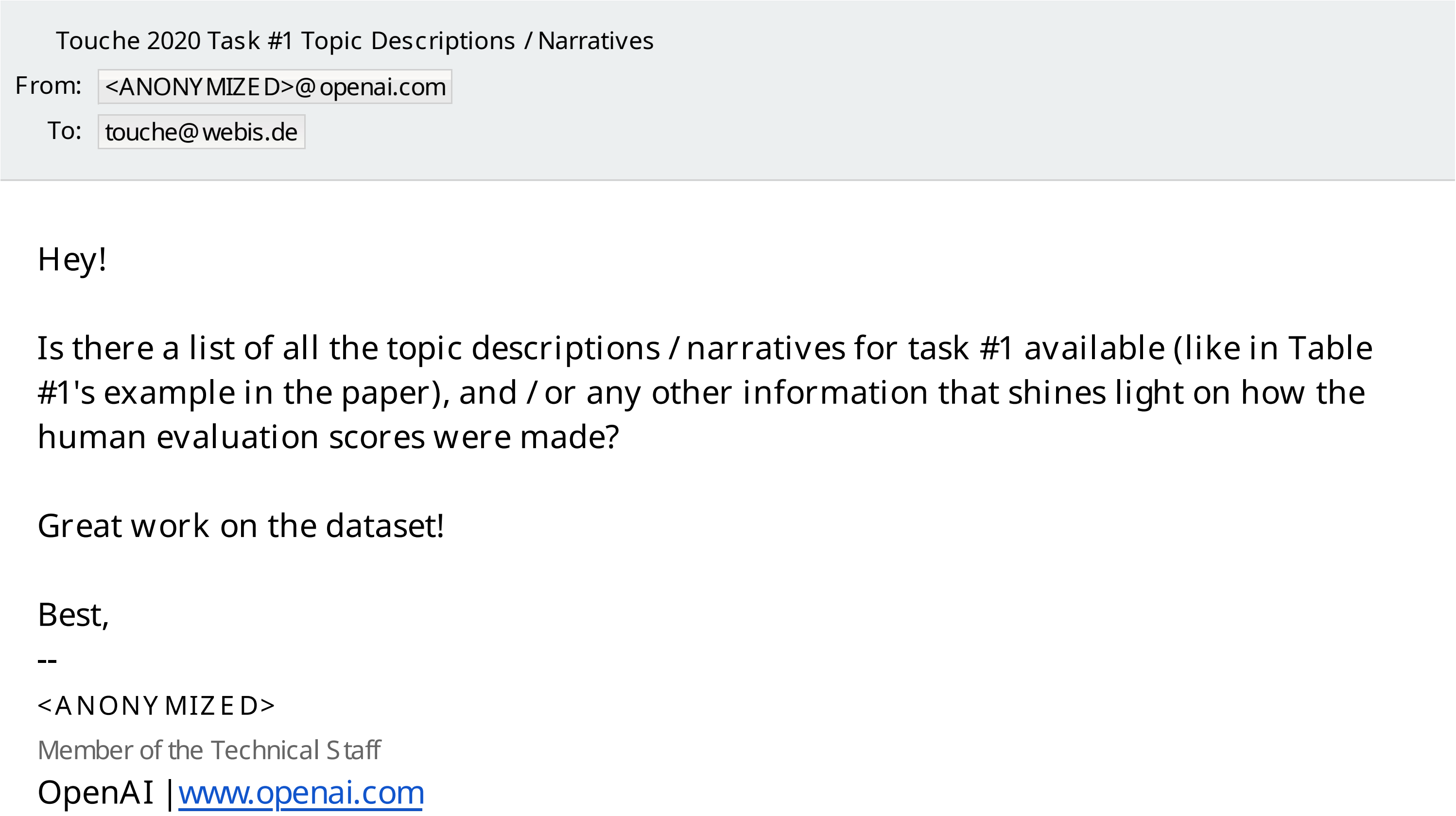


Percy Liang
@percyliang

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.

TIREx: The Information Retrieval Experiment Platform

Problem 3: Blinded Experimentation with LLMs



Touche 2020 Task #1 Topic Descriptions / Narratives

From: <ANONYMIZED>@openai.com

To: touche@webis.de

Hey!

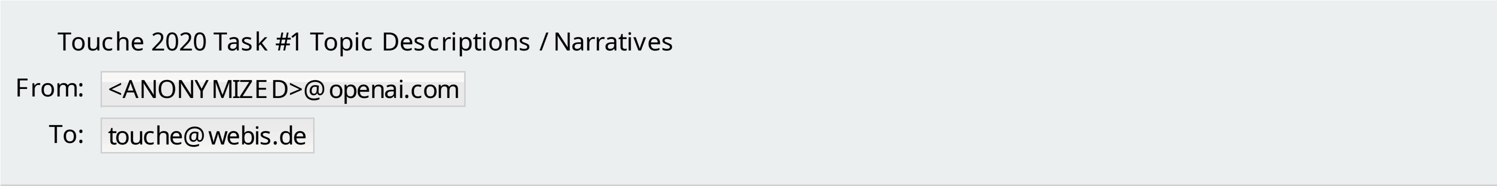
Is there a list of all the topic descriptions / narratives for task #1 available (like in Table #1's example in the paper), and / or any other information that shines light on how the human evaluation scores were made?

Great work on the dataset!

Best,
--
<ANONYMIZED>
Member of the Technical Staff
OpenAI | www.openai.com

TIREx: The Information Retrieval Experiment Platform

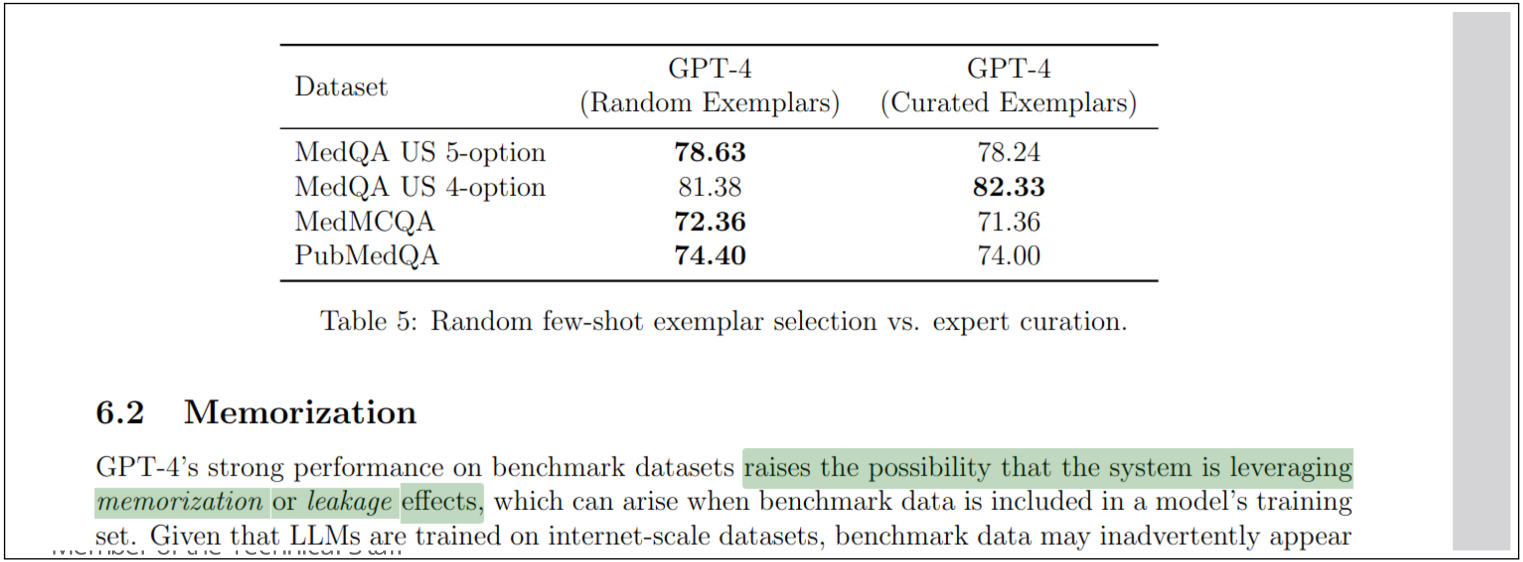
Problem 3: Blinded Experimentation with LLMs



Touche 2020 Task #1 Topic Descriptions / Narratives

From: <ANONYMIZED>@openai.com

To: touche@webis.de



Dataset	GPT-4 (Random Exemplars)	GPT-4 (Curated Exemplars)
MedQA US 5-option	**78.63**	78.24
MedQA US 4-option	81.38	**82.33**
MedMCQA	**72.36**	71.36
PubMedQA	**74.40**	74.00

Table 5: Random few-shot exemplar selection vs. expert curation.

6.2 Memorization

GPT-4's strong performance on benchmark datasets raises the possibility that the system is leveraging *memorization* or *leakage* effects, which can arise when benchmark data is included in a model's training set. Given that LLMs are trained on internet-scale datasets, benchmark data may inadvertently appear

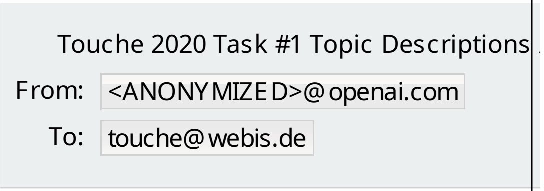
Member of the Technical Staff

OpenAI | www.openai.com

©Webis 2023

TIREx: The Information Retrieval Experiment Platform

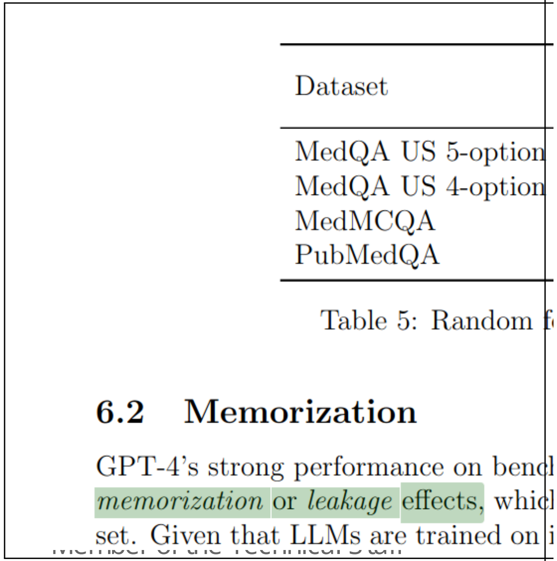
Problem 3: Blinded Experimentation with LLMs



Touche 2020 Task #1 Topic Descriptions

From: <ANONYMIZED>@openai.com

To: touche@webis.de



Dataset
MedQA US 5-option
MedQA US 4-option
MedMCQA
PubMedQA

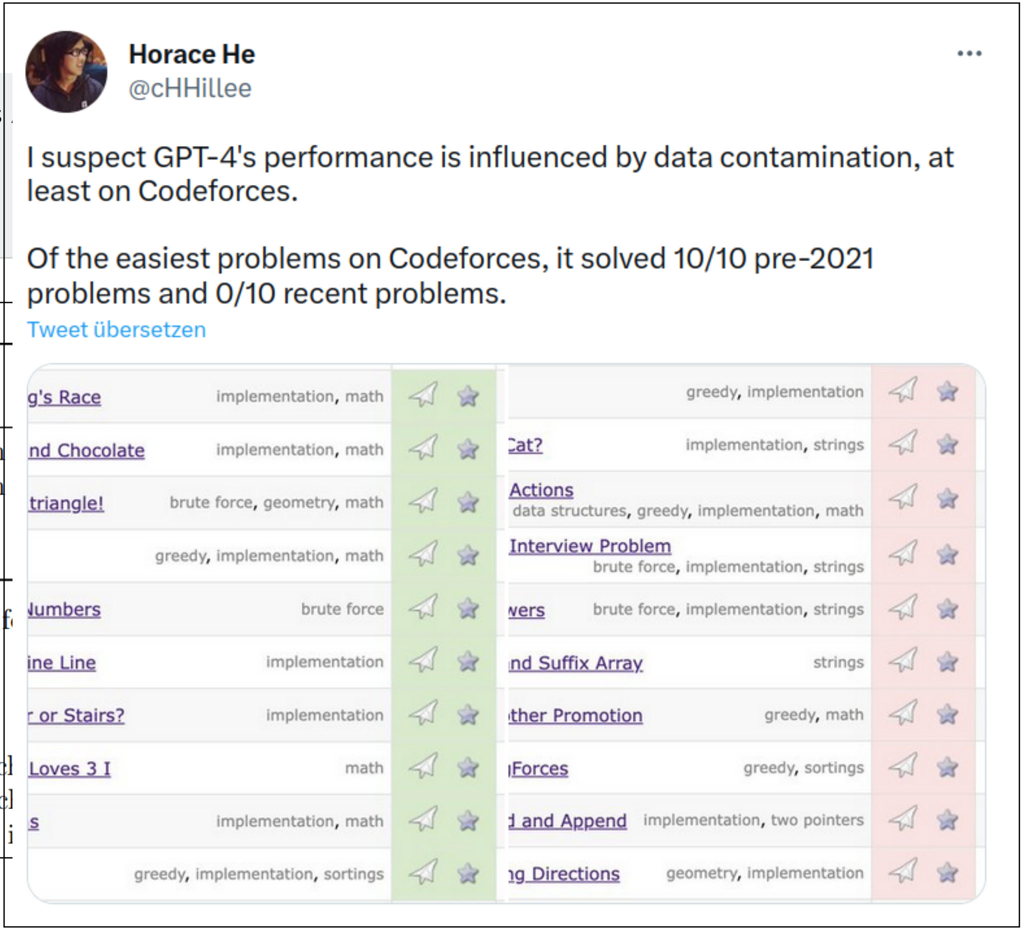
Table 5: Random f

6.2 Memorization

GPT-4's strong performance on bencl *memorization* or *leakage* effects, whicl set. Given that LLMs are trained on i

Member of the Technical Staff

OpenAI | www.openai.com



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.

Tweet übersetzen

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

We Have so Many Tools!

We Have so Many Tools!

Why is Internal and External Validity still a Problem?

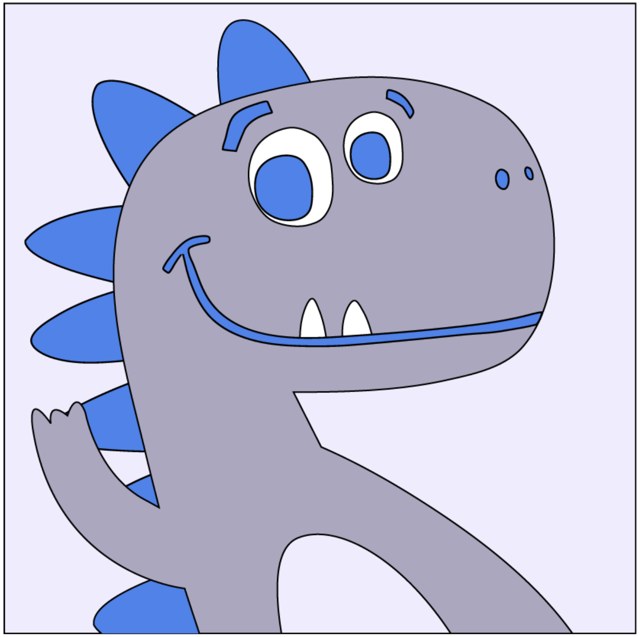
© Webis 2023

We Have so Many Tools!

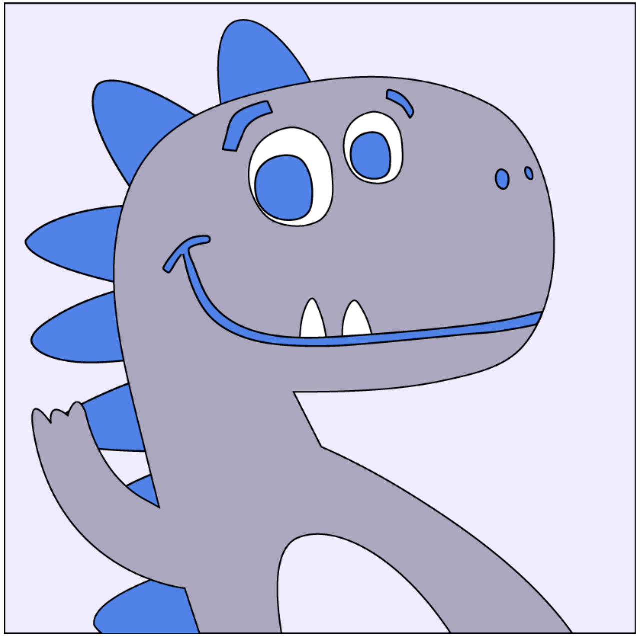
Why is Internal and External Validity still a Problem?

My Hypothesis: Unfavorable Benefit/Effort Ratio

TIREx to the Rescue?



TIREx to the Rescue?



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

© Webis 2023

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

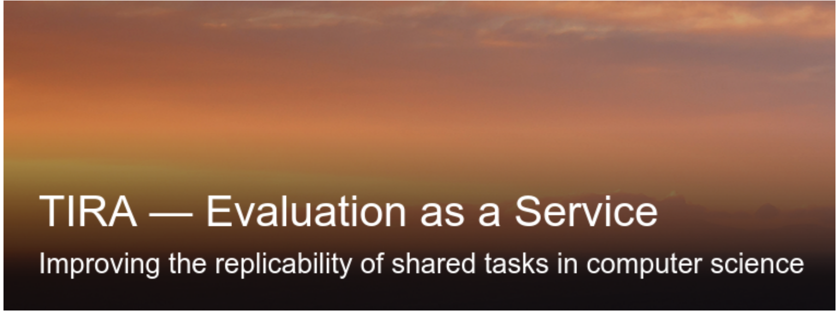
[Gollub’12,Potthast’19,Fröbe’23]

- 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

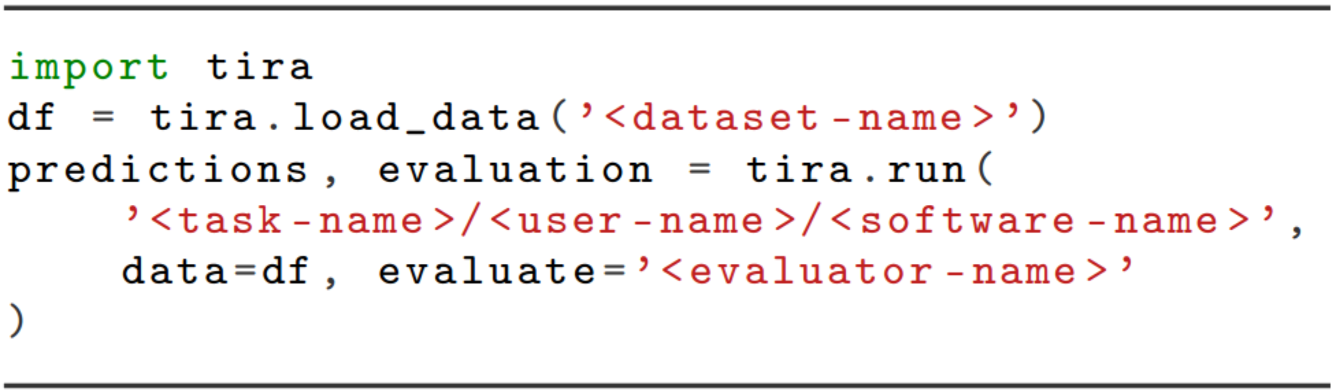
Blinded Experimentation

- 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>'
)
```

- SemEval’23: 2 tasks, 83 + 91 reg. teams (active: 31 + 42; Docker: 21 + 7)
- Enables creative reuse/hacking: `https://values.args.me/`

Enough Preliminaries...

Enough Preliminaries...

Time to get our hands dirty :)

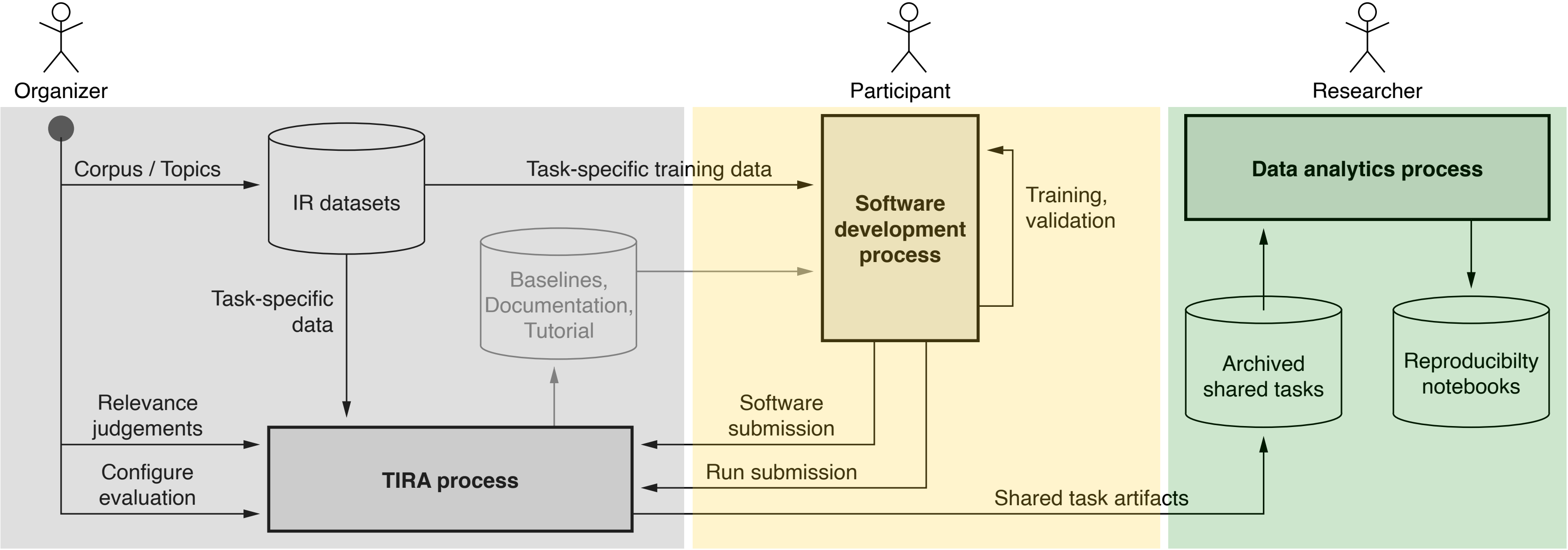
github.com/tira-io/ir-experiment-platform

TIREx: Recap

The hands-on session created two artifacts

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

Covers a shared task end-to-end



© Webis 2023

TIREx: Additional Features

Multi-stage pipelines are first-class citizens

- ❑ Output of previous stages as additional input
- ❑ Caching enabled by immutability of software
  - – Serve output of previous stage if already executed on the dataset

TIREx: Additional Features

Multi-stage pipelines are first-class citizens

- ❑ Output of previous stages as additional input
- ❑ Caching enabled by immutability of software
  - – Serve output of previous stage if already executed on the dataset

Support for external APIs / manual annotations via data uploads

- ❑ Procedure:
  1. Get test inputs from organizers
  2. Produce predictions
  3. Upload predictions
  4. Specify upload as previous stage

TIREx: Additional Features

Multi-stage pipelines are first-class citizens

- ❑ Output of previous stages as additional input
- ❑ Caching enabled by immutability of software
  - – Serve output of previous stage if already executed on the dataset

Support for external APIs / manual annotations via data uploads

- ❑ Procedure:
  1. Get test inputs from organizers
  2. Produce predictions
  3. Upload predictions
  4. Specify upload as previous stage

Support for Re-Rankers

- ❑ Unified data interface via ir_datasets
  - – Allows modularization: Chain arbitrary re-rankers

TIREx: Additional Features

Multi-stage pipelines are first-class citizens

- Output of previous stages as additional input
- Caching enabled by immutability of software
  - Serve output of previous stage if already executed on the dataset

Support for external APIs / manual annotations via data uploads

- Procedure:
  1. Get test inputs from organizers
  2. Produce predictions
  3. Upload predictions
  4. Specify upload as previous stage

Support for Re-Rankers

- Unified data interface via ir_datasets
  - Allows modularization: Chain arbitrary re-rankers

No Lock-in effect

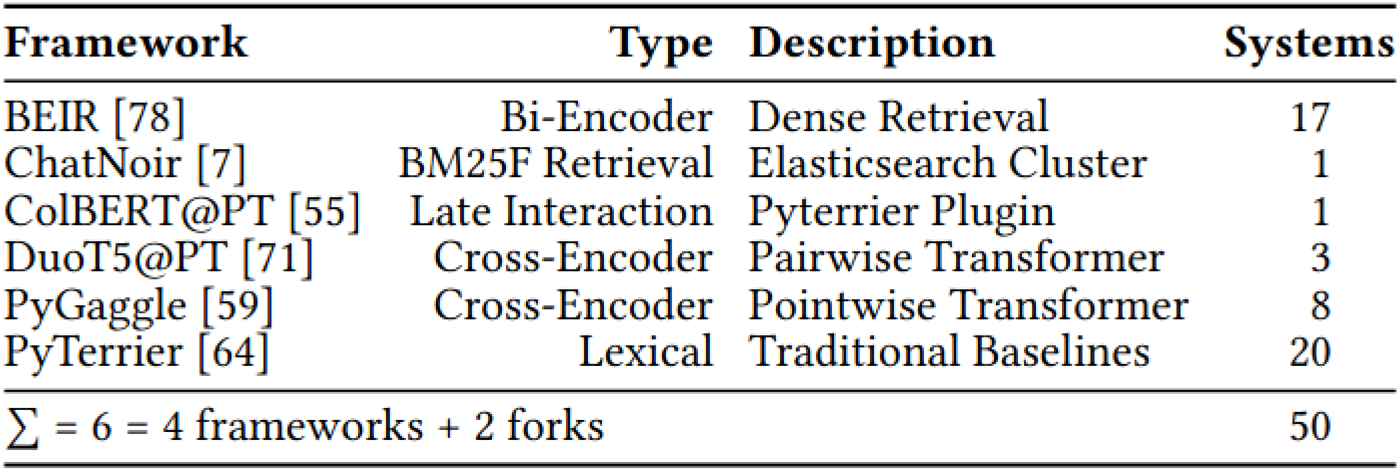
- Example: touche.webis.de/semeval23/touche23-web/tira-software

© Webis 2023

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

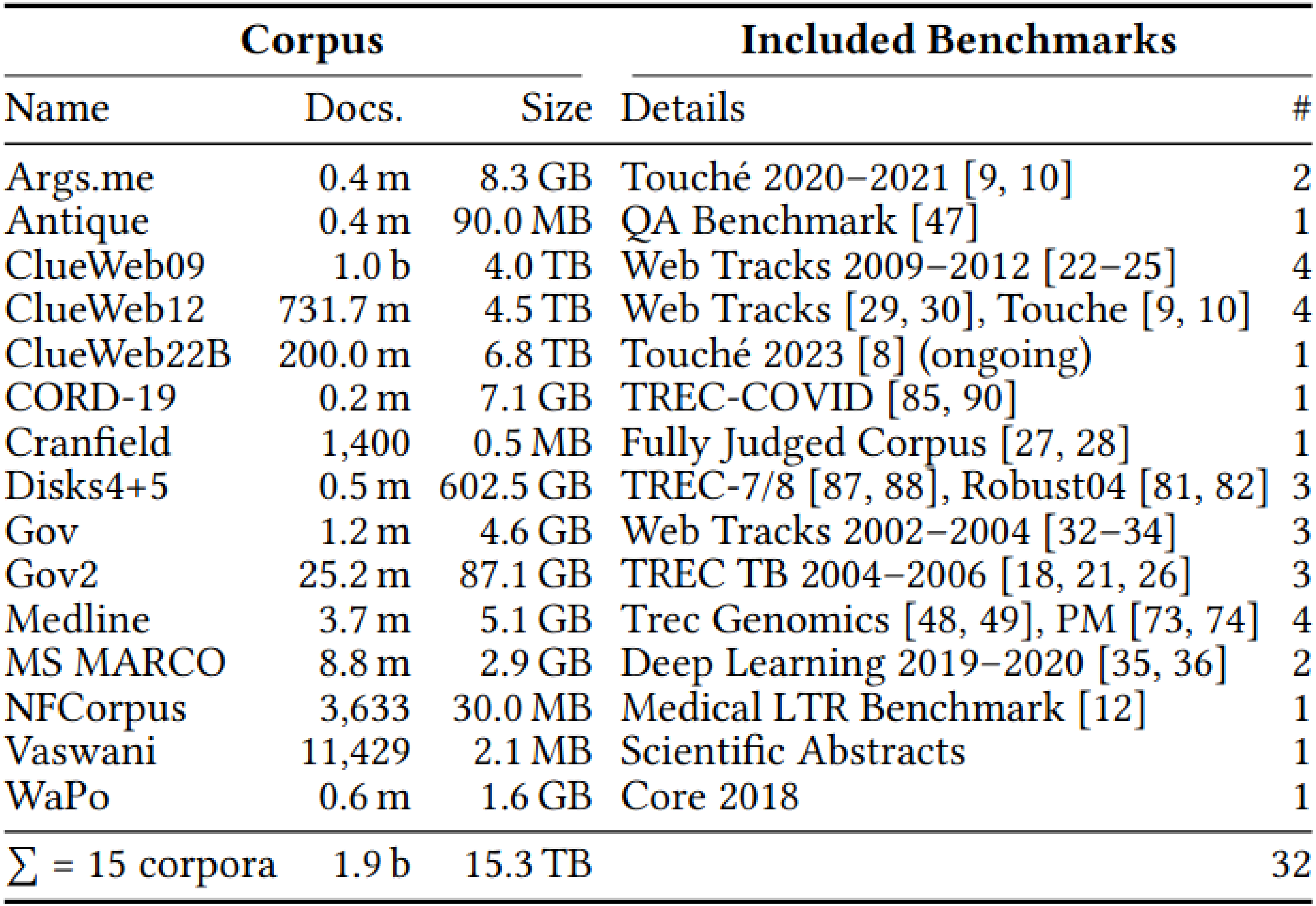


Framework	Type	Description	Systems
BEIR [78]	Bi-Encoder	Dense Retrieval	17
ChatNoir [7]	BM25F Retrieval	Elasticsearch Cluster	1
ColBERT@PT [55]	Late Interaction	Pyterrier Plugin	1
DuoT5@PT [71]	Cross-Encoder	Pairwise Transformer	3
PyGaggle [59]	Cross-Encoder	Pointwise Transformer	8
PyTerrier [64]	Lexical	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
∑ = 15 corpora	1.9 b	15.3 TB		32

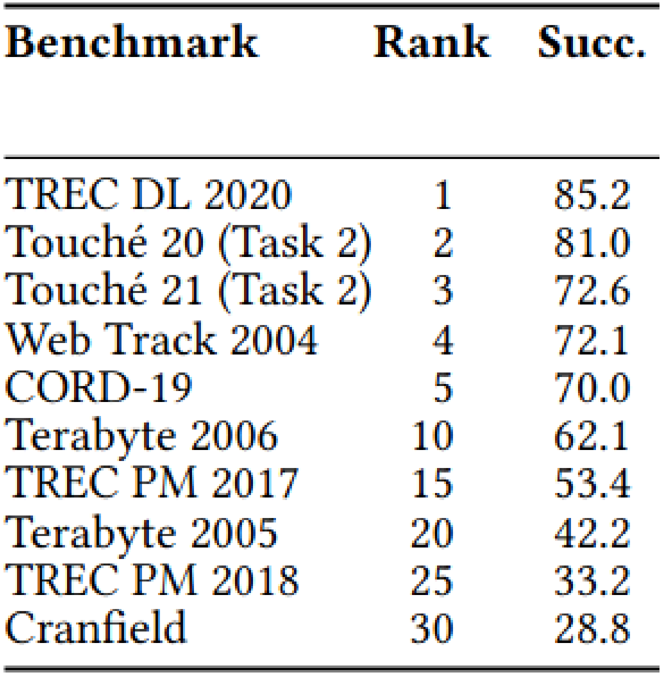
TIREx: Feasibility Study

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

TIREx: Conclusion

Integration of existing tools

- TIRA, ir_datasets, PyTerrier

Better benefit/effort ratio?

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

TIREx: Conclusion

Integration of existing tools

- TIRA, ir_datasets, PyTerrier

Better benefit/effort ratio?

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

Future Work / Wild Guessing

- Move to generative IR (integration of Alpaca?)
- Integration to OWS
  - Link all OWS artifacts to its evaluation in TIREx
  - Three shared tasks are in the setup phase
    1. Index partitioning with selective search
    2. Web genre classification for web crawlers
    3. Spam classification for web crawlers

© Webis 2023

TIREx: Conclusion

Integration of existing tools

- TIRA, ir_datasets, PyTerrier

Better benefit/effort ratio?

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

Future Work / Wild Guessing

- Move to generative IR (integration of Alpaca?)
- Integration to OWS
  - Link all OWS artifacts to its evaluation in TIREx
  - Three shared tasks are in the setup phase
    1. Index partitioning with selective search
    2. Web genre classification for web crawlers
    3. Spam classification for web crawlers

Please Star/Fork

github.com/tira-io/tira

©Webis 2023

TIREx: Conclusion

Integration of existing tools

- TIRA, ir_datasets, PyTerrier

Better benefit/effort ratio?

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

Future Work / Wild Guessing

- Move to generative IR (integration of Alpaca?)
- Integration to OWS
  - Link all OWS artifacts to its evaluation in TIREx
  - Three shared tasks are in the setup phase
    1. Index partitioning with selective search
    2. Web genre classification for web crawlers
    3. Spam classification for web crawlers

Please Star/Fork

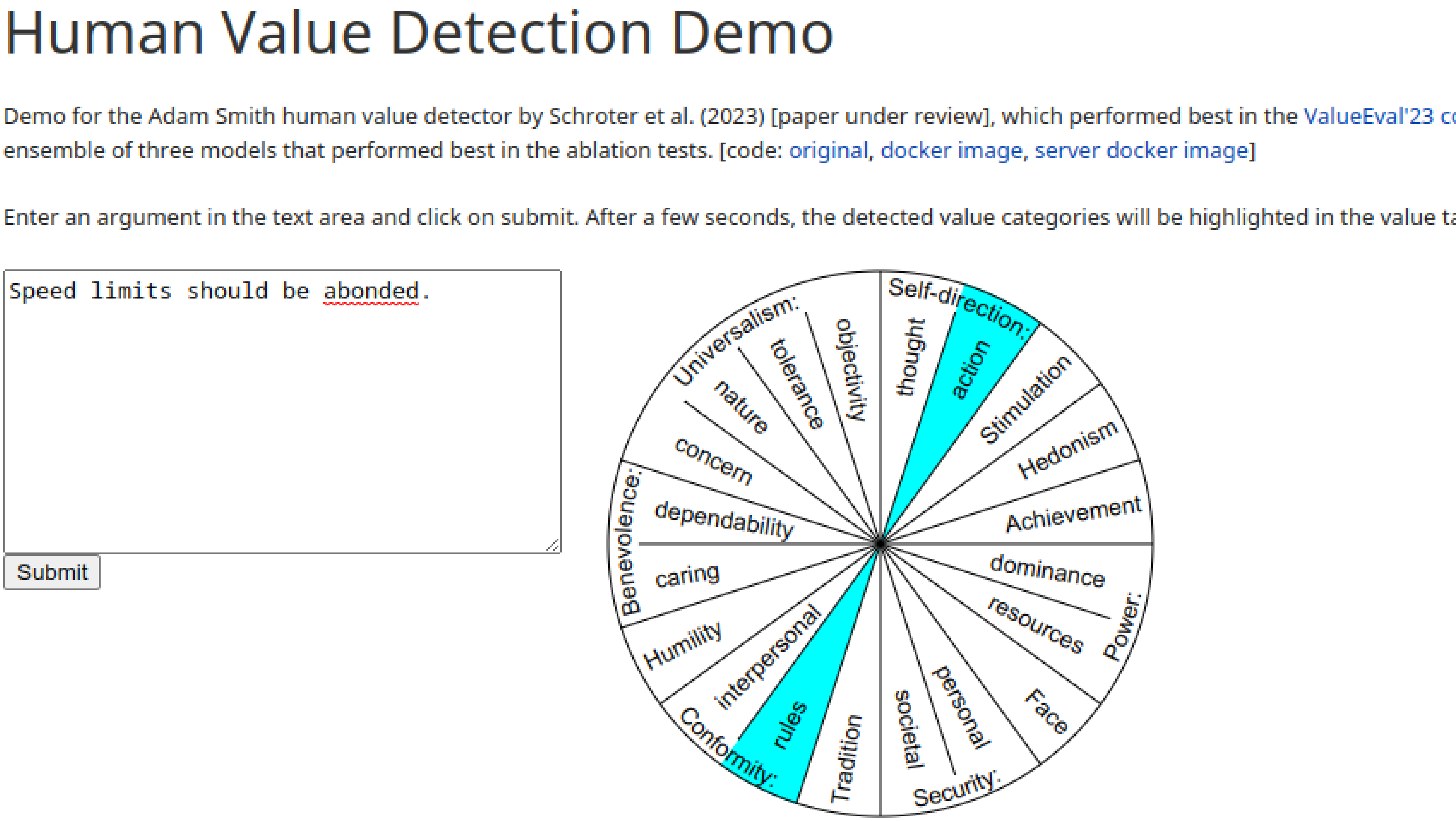
github.com/tira-io/tira

Thank You!

© Webis 2023

© Webis 2023

Backup: SemEval’23 ValueEval Demo (1)



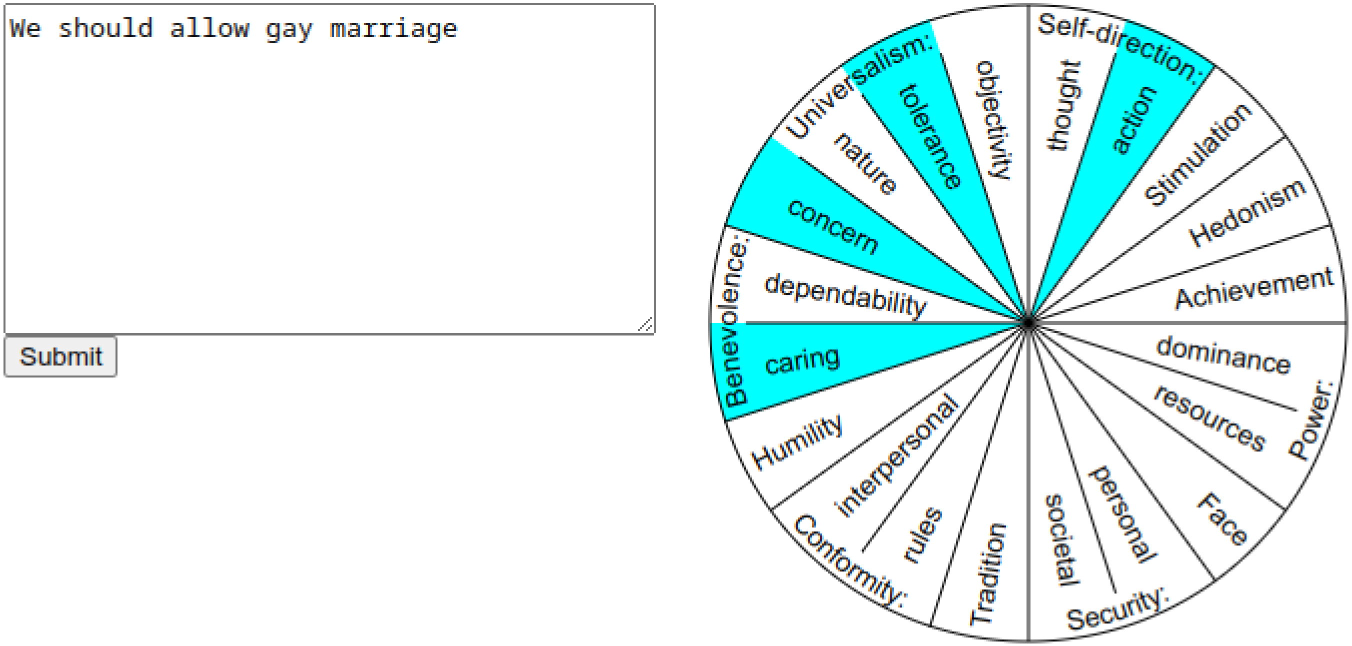
Human Value Detection Demo

Demo for the Adam Smith human value detector by Schroter et al. (2023) [paper under review], which performed best in the ValueEval'23 c
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

© Webis 2023

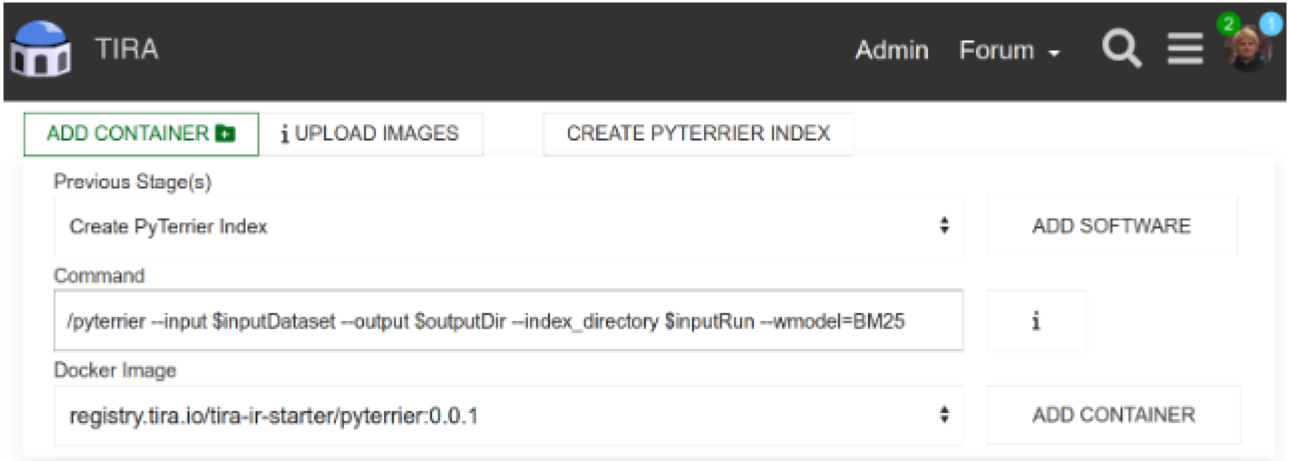
Backup: Limitations

- ❑ 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:

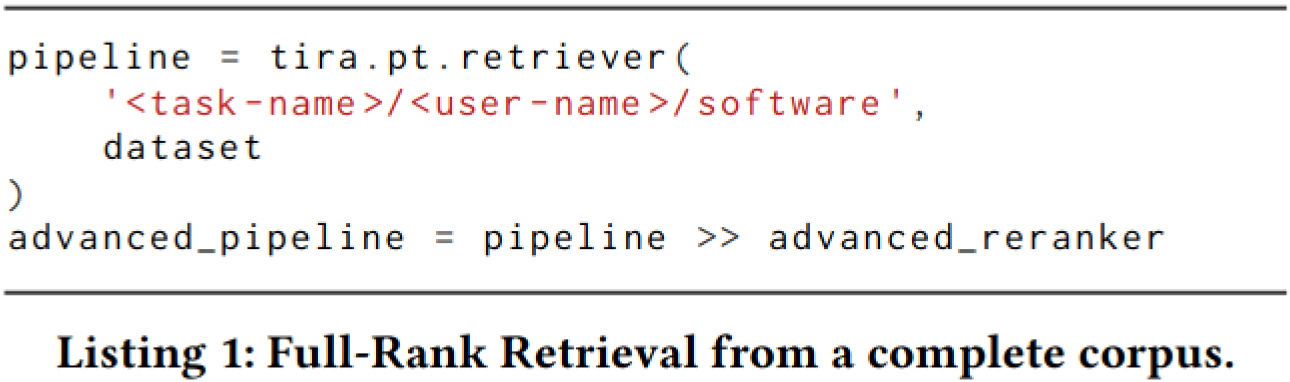
```
/pyterrier --input $inputDataset --output $outputDir --index_directory $inputRun --wmodel=BM25
```

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.

© Webis 2023

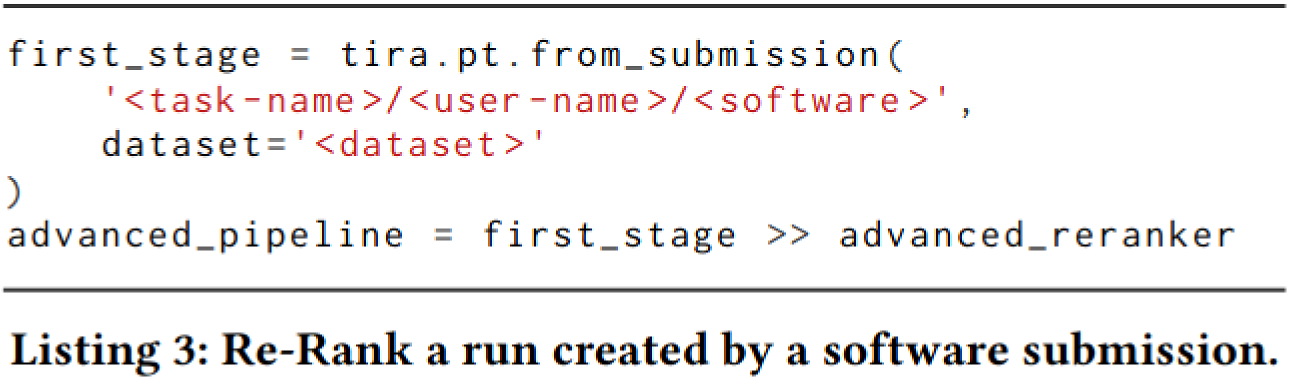
Backup: Full-Rank



```
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.