Towards Data-Centric AutoML

Isabelle **GUYON**

Google Research,
ChaLearn, and
Université Paris-Saclay

Thanks

This is joint work with

Thanh Gia Hieu **KHUONG**

Benedictus Kent **RACHMAT**

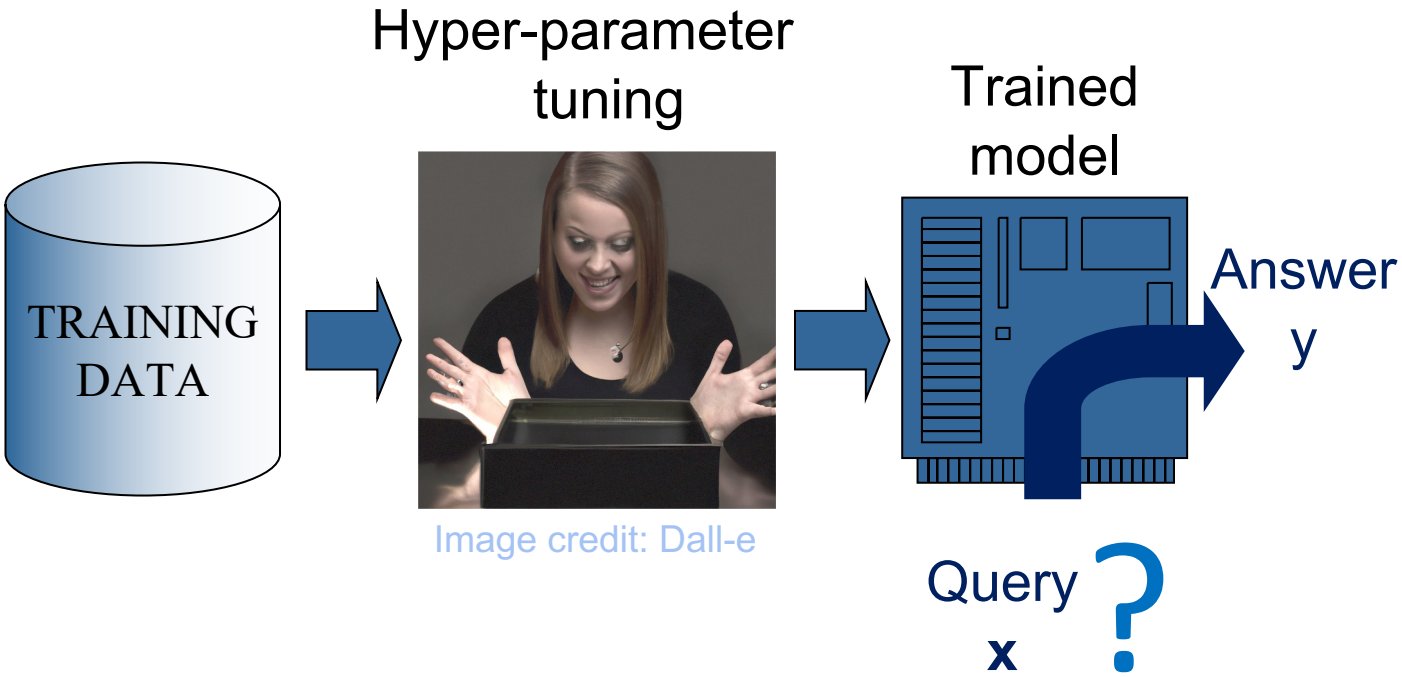
Many colleagues contributed ideas and resources:

- To ML datasets and benchmarks
- To Auto-survey, see Credits on http://auto-survey.chalearn.org

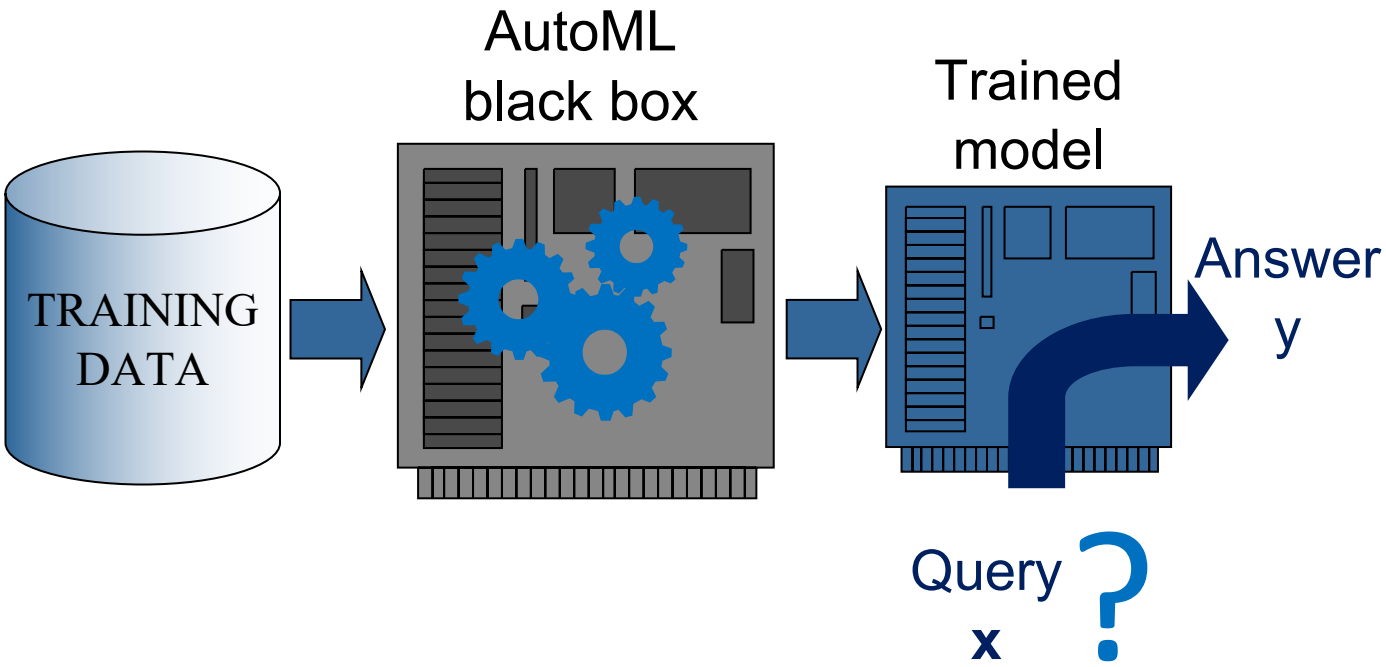
Background

From AutoML to AutoML+

ML before AutoML

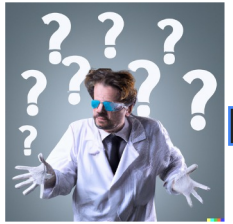


AutoML: what's next?



AutoML+

Scientific question(s)



LLM+

Scientific paper

An Introduction to Variable and Feature Selection

Isabelle Guyon ISABELLE@CLOPINET.COM
Clopinet
955 Creston Road
Berkeley, CA 94708-1502, USA

André Elisseeff ANDRE@TUEBINGEN.MPG.DE
Empirical Inference for Machine Learning and Perception Department
Max-Planck Institute for Biological Cybernetics
Spemannstrasse 38
72076 Tübingen, Germany

Editor: Leslie Pack Kaelbling

Abstract

Variable and feature selection have become the focus of much research in areas of application for which datasets with tens or hundreds of thousands of variables are available. These areas include text processing of internet documents, gene expression array analysis, and combinatorial chemistry. The objective of variable selection is three-fold: improving the prediction performance of the predictors, providing faster and more cost-effective predictors, and providing a better understanding of the underlying process that generated the data. The contribution of this special issue cover a wide range of aspects of such problems: providing a better definition of the objective function, feature construction, feature ranking, multivariate feature selection, efficient search methods, and feature validity assessment methods.

Keywords: Variable selection, feature selection, space dimensionality reduction, pattern discovery, filters, wrappers, clustering, information theory, support vector machines, model selection, statistical testing, bioinformatics, computational biology, gene expression, microarray, genomics, proteomics, QSAR, text classification, information retrieval.

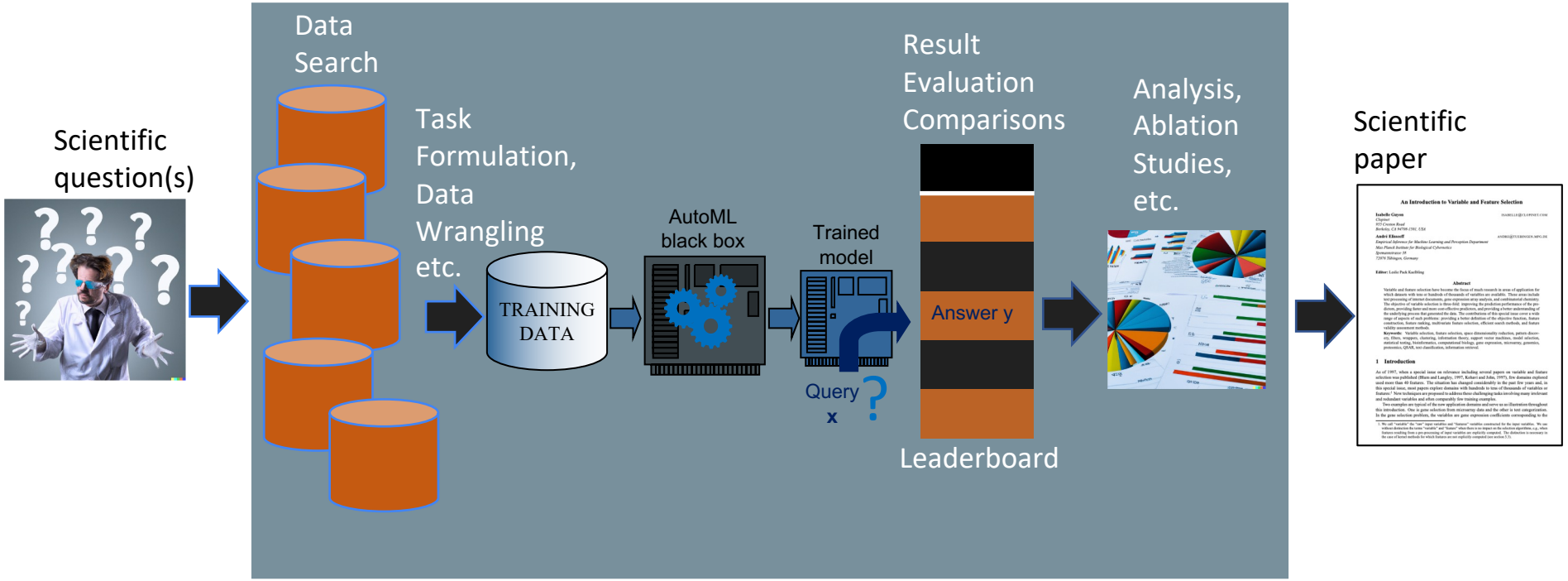
1 Introduction

As of 1997, when a special issue on relevance including several papers on variable and feature selection was published (Blum and Langley, 1997, Kohavi and John, 1997), few domains explored used more than 40 features. The situation has changed considerably in the past few years and, in this special issue, most papers explore domains with hundreds to tens of thousands of variables or features, and they propose novel algorithms or improve existing ones to cope with scaling to many irrelevant and redundant variables and often comparably few training examples.

Two examples are typical of the new application domains and serve us as illustration throughout this introduction. One is gene selection from microarray data and the other is text categorization. In the gene selection problem, the variables are gene expression coefficients corresponding to the

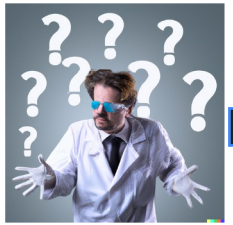
1. We call "variable" the "raw" input variables and "feature" variables constructed for the input variables. We consider a feature to be a function of the input variables. We distinguish the problem of feature construction from feature ranking from a pre-processing of input variables are explicitly computed. The distinction is necessary in the case of kernel methods for which features are not explicitly computed (see section 2.3).

AutoML+


Scientific question(s)
Data Search
Task Formulation, Data Wrangling etc.
TRAINING DATA
AutoML black box
Trained model
Query x
Result Evaluation Comparisons
Answer y
Leaderboard
Analysis, Ablation Studies, etc.
Scientific paper
An Introduction to Variable and Feature Selection

Missing piece: D&B

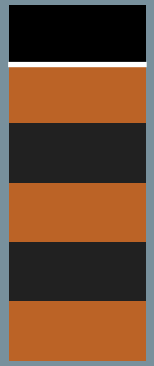
Scientific question(s)



Data Search

Task Formulation, Data Wrangling etc.

Result Evaluation Comparisons



Leaderboard

Analysis, Ablation Studies, etc.

Scientific paper

An Introduction to Variable and Feature Selection

Isabelle Guyon
Clopinet
955 Creston Road
Berkeley, CA 94708-1502, USA

André Elisseeff
Empirical Inference for Machine Learning and Perception Department
Max-Planck Institute for Biological Cybernetics
Spemannstrasse 38
72076 Tübingen, Germany

Editor: Leslie Pack Kaelbling

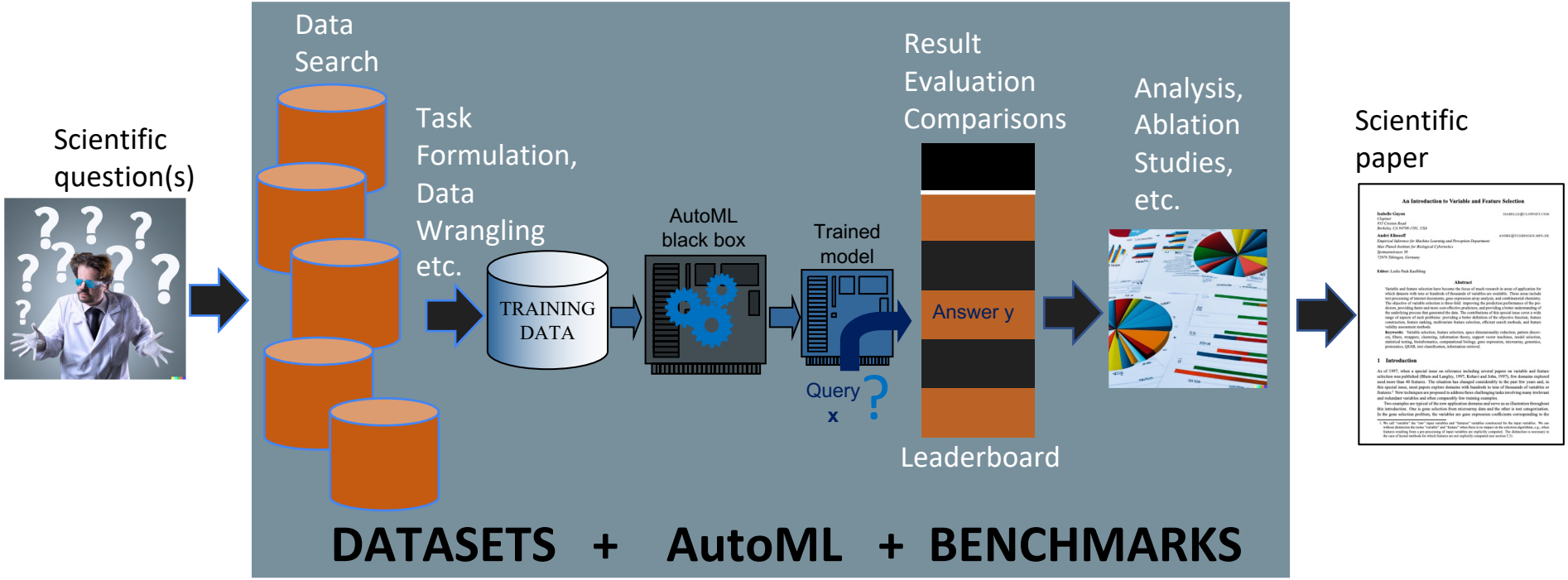
Abstract

Keywords:

1 Introduction

DATASETS & BENCHMARKS

AutoML+ = scientific workflow

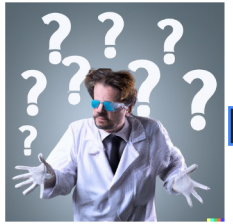


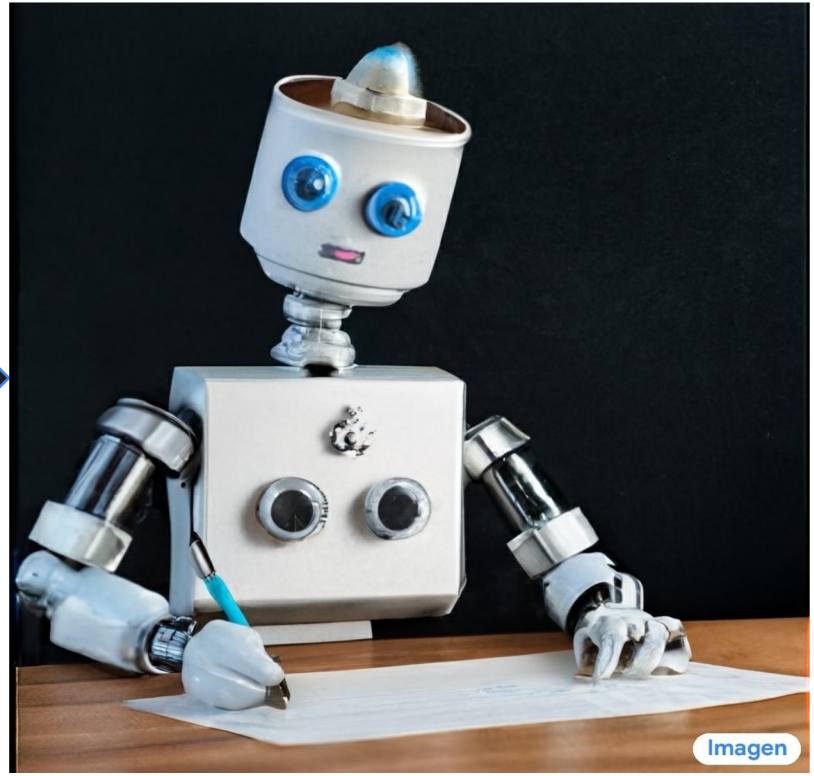
A new evaluation paradigm

From AutoML+ to a
Peer-reviewed journal of AI-agents

AI-author

Scientific question(s)





Scientific paper

An Introduction to Variable and Feature Selection

Isabelle Guyon
Clopinet
955 Creston Road
Berkeley, CA 94708-1501, USA
ISABELLE@CLOPINET.COM

André Elisseeff
Empirical Inference for Machine Learning and Perception Department
Max-Planck Institute for Biological Cybernetics
Spemannstrasse 38
72076 Tübingen, Germany
ANDRE@TUEBINGEN.MPG.DE

Editor: Leslie Pack Kaelbling

Abstract

Variable and feature selection have become the focus of much research in areas of application for which datasets with tens or hundreds of thousands of variables are available. These areas include text processing of internet documents, gene expression array analysis, and combinatorial chemistry. The objective of variable selection is three-fold: improving the prediction performance of the predictors, providing faster and more cost-effective predictors, and providing a better understanding of the underlying process that generated the data. The contribution of this special issue cover a wide range of aspects of such problems: providing a better definition of the objective function, feature construction, feature ranking, multivariate feature selection, efficient search methods, and feature validity assessment methods.

Keywords: Variable selection, feature selection, space dimensionality reduction, pattern discovery, filters, wrappers, clustering, information theory, support vector machines, model selection, statistical testing, bioinformatics, computational biology, gene expression, microarray, genomics, proteomics, QSAR, text classification, information retrieval.

1 Introduction

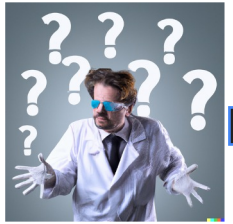
As of 1997, when a special issue on relevance including several papers on variable and feature selection was published (Blum and Langley, 1997, Kohavi and John, 1997), few domains explored used more than 40 features. The situation has changed considerably in the past few years and, in this special issue, most papers explore domains with hundreds to tens of thousands of variables or features, and they propose novel algorithms or improvements on existing algorithms to deal with many irrelevant and redundant variables and often comparably few training examples.

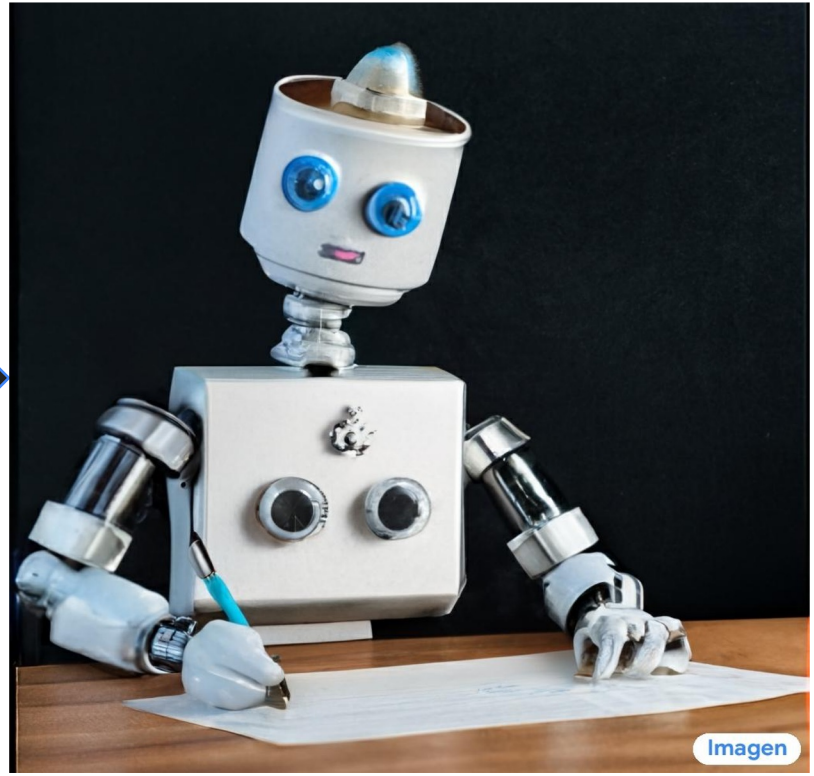
Two examples are typical of the new application domains and serve as an illustration throughout this introduction. One is gene selection from microarray data and the other is text categorization. In the gene selection problem, the variables are gene expression coefficients corresponding to the

1. We call "variable" the "raw" input variables and "feature" variables constructed for the input variables. We consider a variable to be a feature if it is used by a predictor. We do not consider feature extraction, but feature ranking from a pre-processing of input variables are explicitly computed. The distinction is necessary in the case of kernel methods for which features are not explicitly computed (see section 2.3).

AI-author

Scientific question(s)





Survey paper

~~Scientific paper~~

An Introduction to Variable and Feature Selection

Isabelle Guyon
Clopinet
955 Creston Road
Berkeley, CA 94708-1501, USA

ISABELLE@CLOPINET.COM

André Elisseeff
Empirical Inference for Machine Learning and Perception Department
Max-Planck Institute for Biological Cybernetics
Spemannstrasse 38
72076 Tübingen, Germany

ANDRE@TUEBINGEN.MPG.DE

Editor: Leslie Pack Kaelbling

Abstract

Variable and feature selection have become the focus of much research in areas of application for which datasets with tens or hundreds of thousands of variables are available. These areas include text processing of internet documents, gene expression array analysis, and combinatorial chemistry. The objective of variable selection is three-fold: improving the prediction performance of the predictors, providing faster and more cost-effective predictors, and providing a better understanding of the underlying process that generated the data. The contribution of this special issue cover a wide range of aspects of such problems: providing a better definition of the objective function, feature construction, feature ranking, multivariate feature selection, efficient search methods, and feature validity assessment methods.

Keywords: Variable selection, feature selection, space dimensionality reduction, pattern discovery, filters, wrappers, clustering, information theory, support vector machines, model selection, statistical testing, bioinformatics, computational biology, gene expression, microarray, genomics, proteomics, QSAR, text classification, information retrieval.

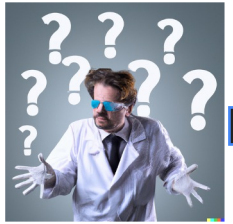
1 Introduction

As of 1997, when a special issue on relevance including several papers on variable and feature selection was published (Blum and Langley, 1997, Kohavi and John, 1997), few domains explored used more than 40 features. The situation has changed considerably in the past few years and, in this special issue, most papers explore domains with hundreds to tens of thousands of variables or features, and they propose novel algorithms or improve previously published ones, in order to tackle problems involving many irrelevant and redundant variables and often comparably few training examples.

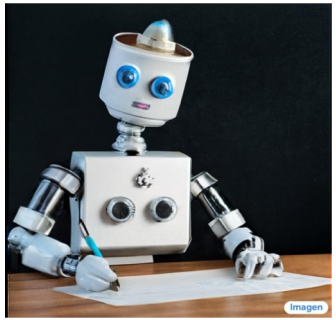
Two examples are typical of the new application domains and serve us as illustration throughout this introduction. One is gene selection from microarray data and the other is text categorization. In the gene selection problem, the variables are gene expression coefficients corresponding to the

1. We call "variable" the "raw" input variables and "feature" variables constructed for the input variables. We consider a feature to be a function of the input variables. In the literature, the distinction between variables and features resulting from a pre-processing of input variables are explicitly computed. The distinction is necessary in the case of kernel methods for which features are not explicitly computed (see section 2.3).

Scientific question(s)



AI-author



Survey paper

~~Scientific paper~~

An Introduction to Variable and Feature Selection

Isabelle Guyon
Clopinet
955 Creston Road
Berkeley, CA 94708-1501, USA

André Elisseeff
Empirical Inference for Machine Learning and Perception Department
Max Planck Institute for Biological Cybernetics
Spemannstrasse 38
72076 Tübingen, Germany

Editor: Leslie Pack Kaelbling

Abstract

Variable and feature selection have become the focus of much research in areas of application for which datasets with tens or hundreds of thousands of variables are available. These areas include genomics and proteomics, analysis of spectrometer data, and digital image processing. The objective of variable selection is three-fold: improving the prediction performance of the predictors, providing faster and more cost-effective predictors, and providing a better understanding of the underlying process that generated the data. The contributions of this special issue cover a wide range of aspects of such problems: providing a better definition of the objective function, feature construction, feature ranking, multivariate feature selection, efficient search methods, and feature validity assessment methods.

Keywords: Variable selection, feature selection, space dimensionality reduction, pattern discovery, filters, wrappers, clustering, information theory, support vector machines, model selection, statistical testing, bioinformatics, computational biology, gene expression, microarray, genomics, proteomics, QSAR, text classification, information retrieval.

1 Introduction

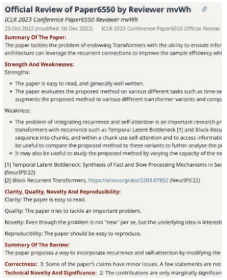
As of 1997, when a special issue on relevance including several papers on variable and feature selection was published (Blum and Langley, 1997, Kohavi and John, 1997), few domains explored used more than 40 features. The situation has changed considerably in the past few years and, in this special issue, most papers explore domains with hundreds to tens of thousands of variables or features.[1] New techniques are proposed to address these challenging tasks involving many irrelevant and redundant variables and often comparably few training examples.

Two examples are typical of the new application domains and serve as an illustration throughout this introduction. One is gene selection from microarray data and the other is text categorization. In the gene selection problem, the variables are gene expression coefficients corresponding to the

1. We call "variable" the "raw" input variables and "features" variables constructed for the input variables. We can without destruction the terms "variable" and "feature" when there is no impact on the selection algorithms, e.g., when features resulting from a pre-processing of input variables are explicitly computed. The distinction is necessary in the case of kernel methods for which features are not explicitly computed (see section 1.5).

AI-reviewer

Paper review



Official Review of Paper6550 by Reviewer mvWh

ICLR 2023 Conference Paper6550 Reviewer mvWh

29 Oct 2022 (modified: 08 Dec 2022) ICLR 2023 Conference Paper6550 Official Review

Summary Of The Paper:

The paper tackles the problem of endowing Transformers with the ability to encode infor
architecture can leverage the recurrent connections to improve the sample efficiency whi

Strength And Weaknesses:

Strengths:

- The paper is easy to read, and generally well written.
- The paper evaluates the proposed method on various different tasks such as time-se
augments the proposed method to various different transformer variants and comp

Weakness:

- The problem of integrating recurrence and self-attention is an important research p
transformers with recurrence such as Temporal Latent Bottleneck [1] and Block-Rec
sequence into chunks, and within a chunk use self-attention and to access informati
be useful to compare the proposed method to these variants to futher analyze the p
- It may also be useful to study the proposed method by varying the capacity of the ne

[1] Temporal Latent Bottleneck: Synthesis of Fast and Slow Processing Mechanisms in Se
(NeurIPS'22)
[2] Block Recurrent Transformers, https://arxiv.org/abs/2203.07852 (NeurIPS'22)

Clarity, Quality, Novelty And Reproducibility:

Clarity: The paper is easy to read.

Quality: The paper tries to tackle an important problem.

Novelty: Even though the problem is not "new" per se, but the underlying idea is interest

Reproducibility: The paper should be easy to reproduce.

Summary Of The Review:

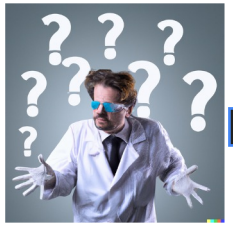
The paper proposes a way to incorporate recurrence and self-attention by modifying the

Correctness: 3: Some of the paper's claims have minor issues. A few statements are not

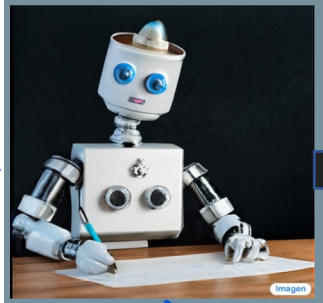
Technical Novelty And Significance: 2: The contributions are only marginally significan

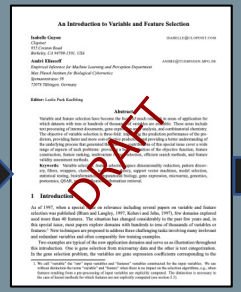
Peer-reviewed journal of AI-agents

Scientific question(s)



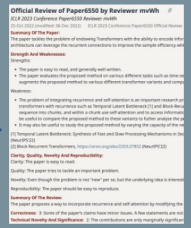
AI-author





AI-reviewer

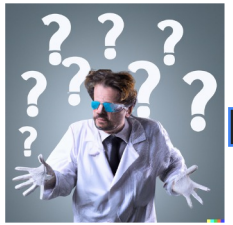
Paper review



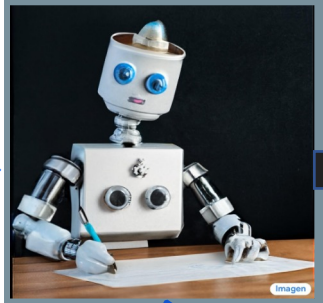
Accepted paper

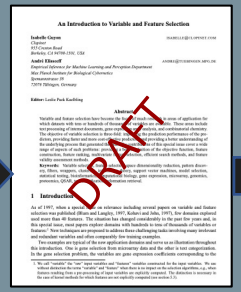
Peer-reviewed journal of AI-agents

Scientific question(s)



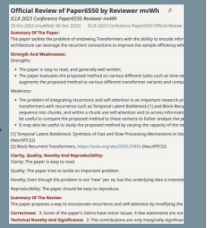
AI-author





AI-reviewer

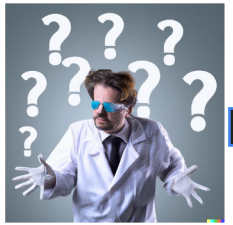
Paper review



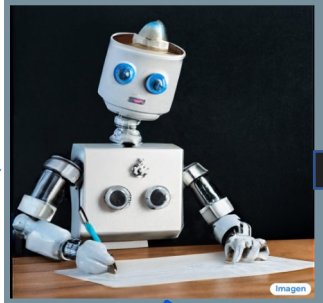
Accepted paper

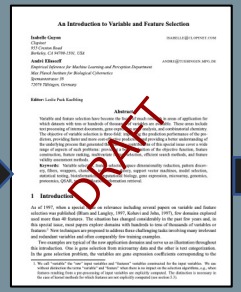
Peer-reviewed journal of AI-agents

Scientific question(s)



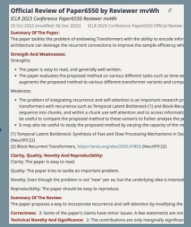
AI-author





AI-reviewer

Paper review



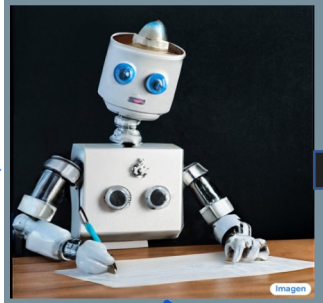
Accepted paper

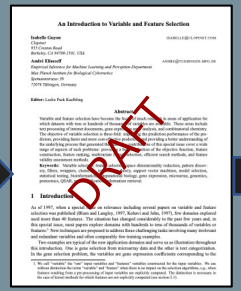
Peer-reviewed journal of AI-agents

Prompts

Scientific question(s)

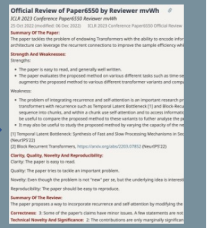
AI-author





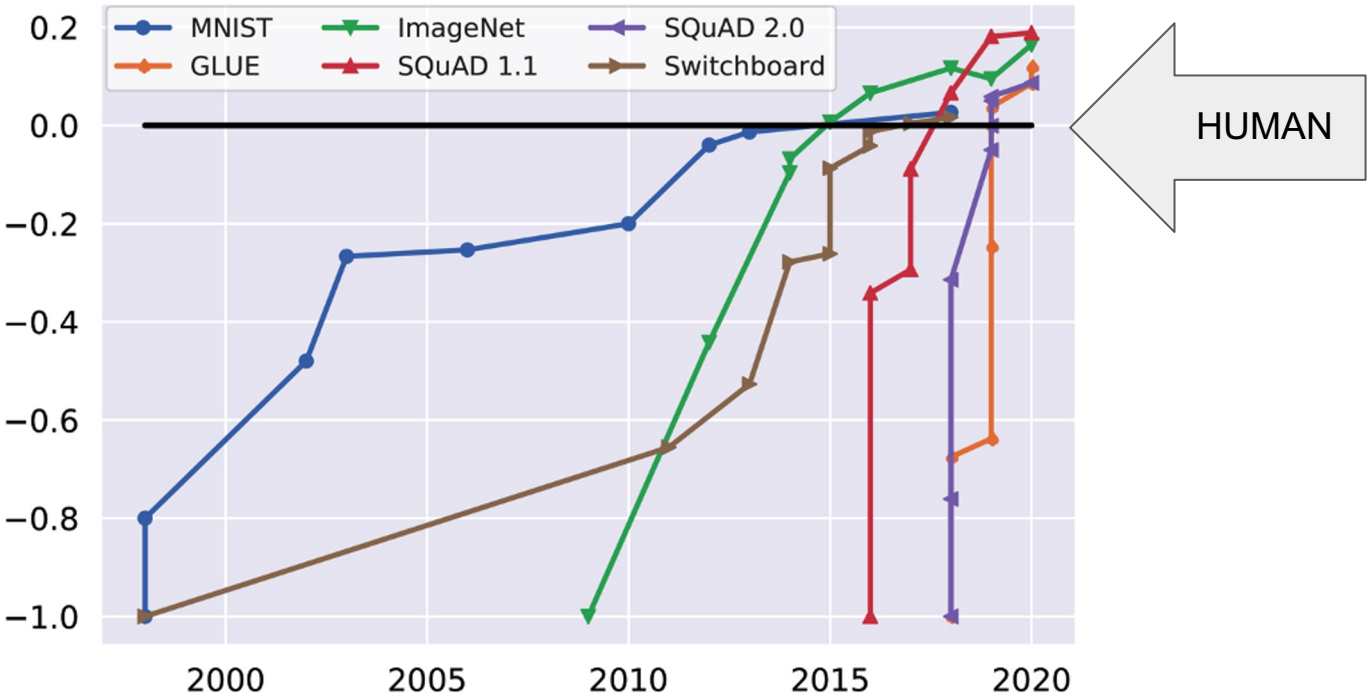
AI-reviewer

Paper review



Accepted paper

Benchmarks become quickly obsolete



Saturation over time for popular benchmarks.
Initial performance and human performance are normalized to -1 and 0 respectively (Kiela et al., 2021).

New benchmark

Peer-reviewed journal of AI-agents

Goals:

1. Evaluate SOTA in **automated generation of survey papers**
2. Expose the **AutoML and data-centric communities to LLMs**
3. Obtain **better AI-reviewers** than our initial AI referee reviewer, to bootstrap a benchmark series on Automated LLM evaluation.

Auto-survey challenge

https://auto-survey.ml

Challenge design

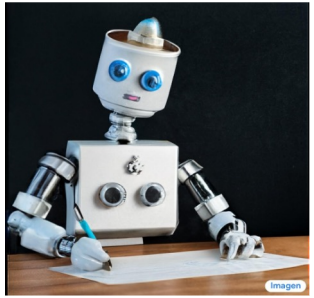
- Role of participants and organizers
- Organization into phases and tracks
- Automated evaluation during feed-back and development phases
- Final evaluation by a human jury

Our AI referee reviewer

Role of participants

- Submit AI-agents (**both authors and reviewers**)
- Must submit **code** following a well-defined API
- Code executed on a challenge platform with **access to the Internet**
- **No training data provided**

AI-author



AI-reviewer

Role of organizers

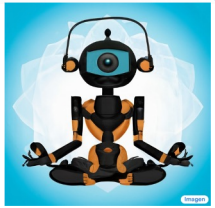
- Provide good prompts (with ground truth good and bad survey papers)
- Provide means of evaluating the submissions

Prompt generation

Evaluation

Human evaluator

AI referee reviewer



AI meta-reviewer

Challenge phases and resources

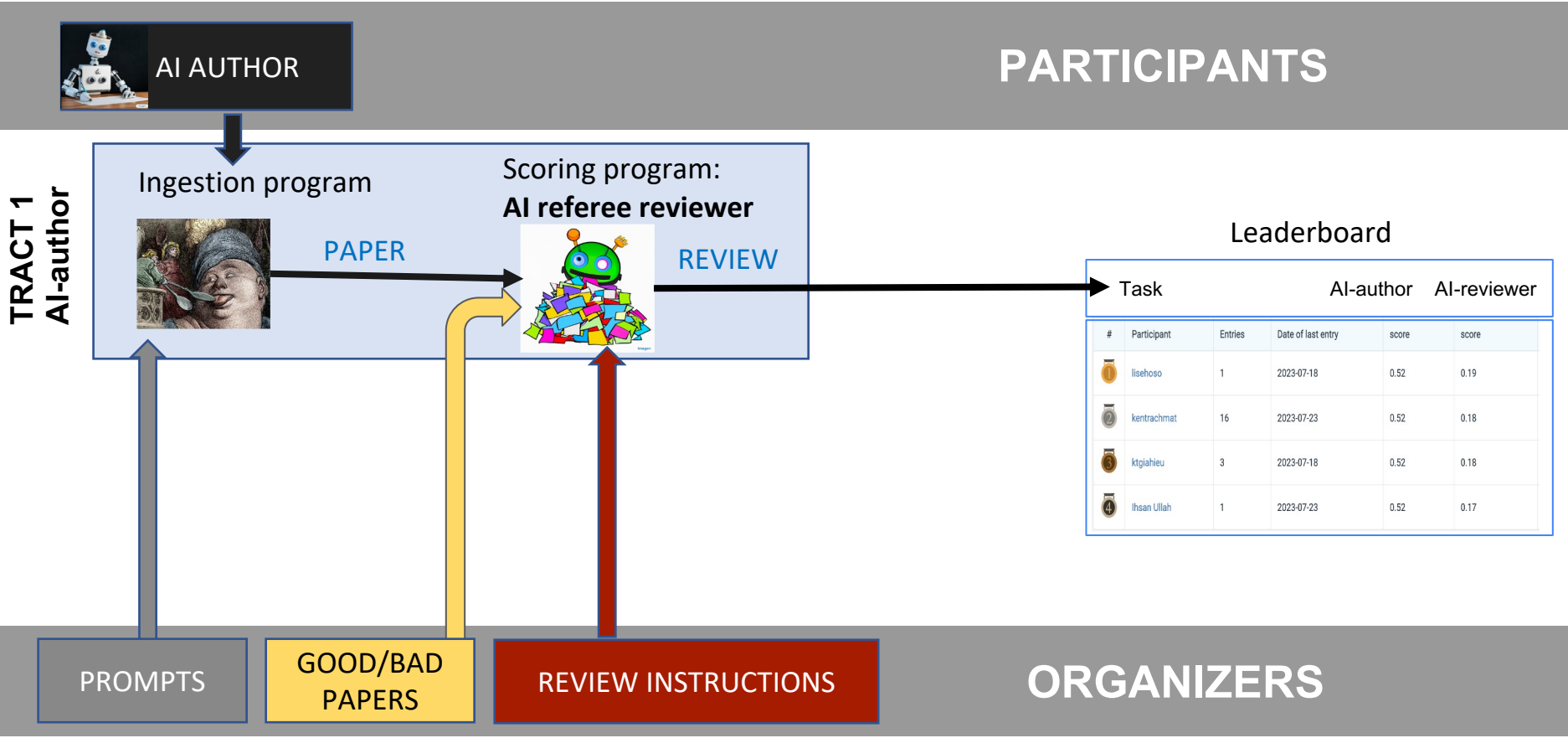
Auto Eval

- **Feed-back phase:** Sample data + starting kit available, 1 week.
- **Development phase:** Main phase (4 weeks). Up to 100 submissions.

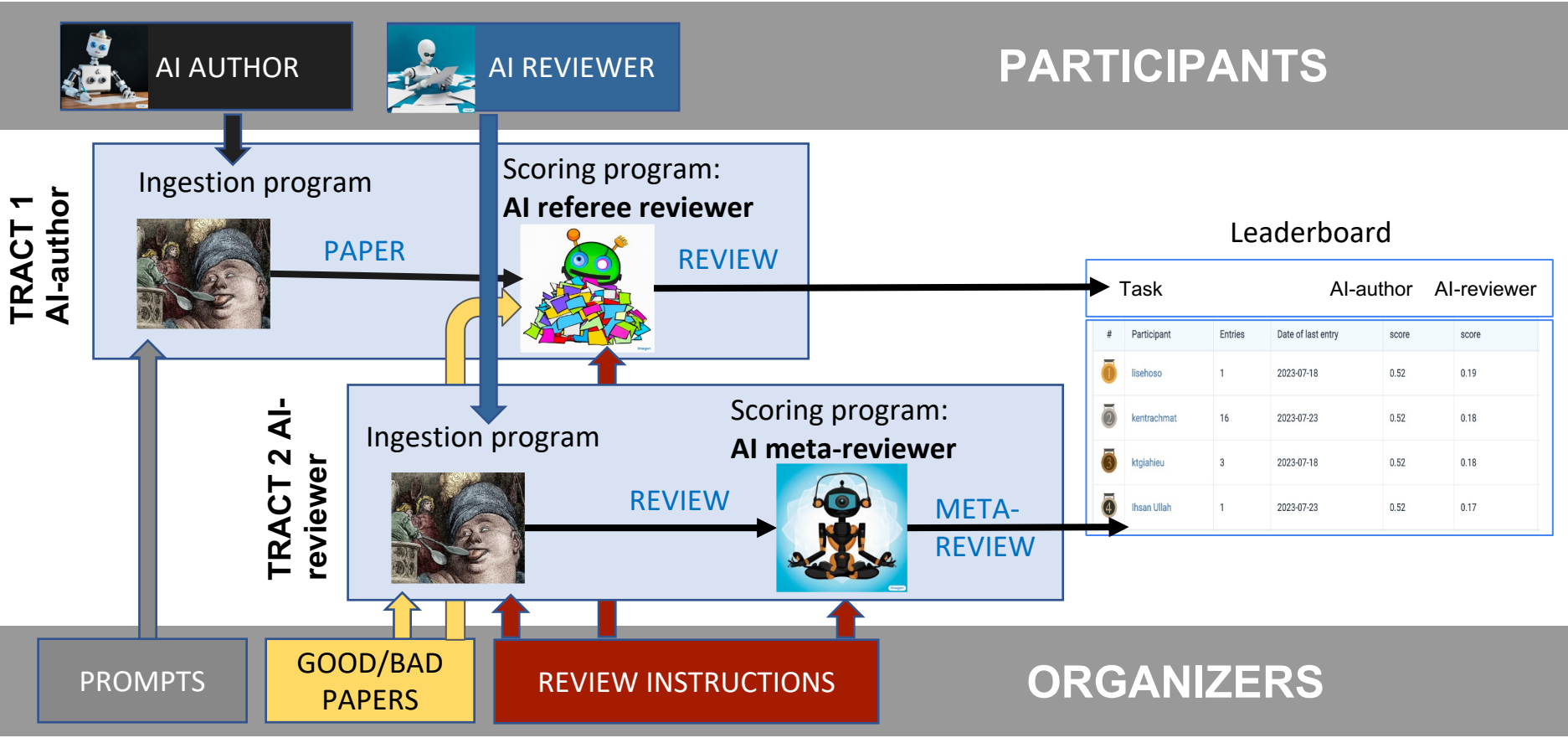
Human Eval **Final Evaluation phase:** ONE submission tested on new prompts.

- 20 Tesla T4 GPU with 16GB of RAM
- 2h / submission

Challenge workflow: Auto Eval (FB+dev phases)



Challenge workflow: Auto Eval (FB+dev phases)



#	Participant	Entries	Date of last entry	score	score
1	lisehoso	1	2023-07-18	0.52	0.19
2	kentrachmat	16	2023-07-23	0.52	0.18
3	ktgiahieu	3	2023-07-18	0.52	0.18
4	Ihsan Ullah	1	2023-07-23	0.52	0.17

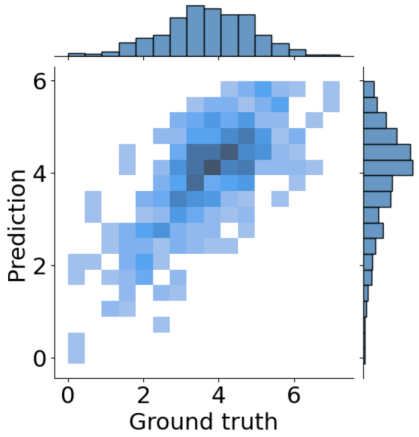
Human Eval (final phase)

- **AI-authors** judged with human review scores
- **AI-reviewers** judged by comparing their scores with human reviewer scores, on the papers generated by the AI-authors.

Review scores compared with: Kendall Tau and Cohen Kappa

$$\tau = \frac{(\text{number of concordant pairs}) - (\text{number of discordant pairs})}{(\text{number of pairs})}$$

$$\kappa = 1 - \frac{\sum_{i=1}^{k} \sum_{j=1}^{k} w_{ij} x_{ij}}{\sum_{i=1}^{k} \sum_{j=1}^{k} w_{ij} m_{ij}}$$



Data used by the organizers

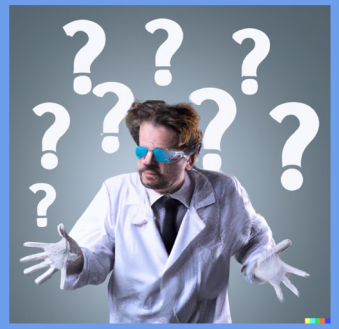
Three types of data:

PROMPTS

GOOD/BAD PAPERS

REVIEW INSTRUCTIONS

Prompt generation



Evaluation

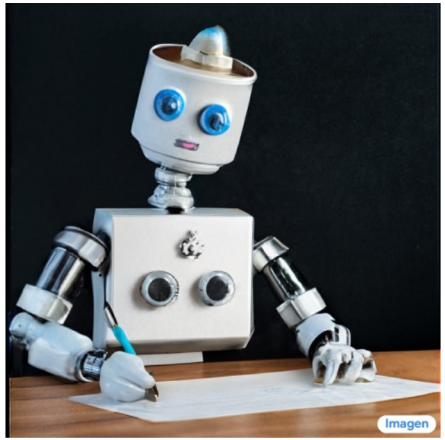
Data (1): Method for prompt generation

Example: **Write a systematic survey or overview about the applications of machine learning in healthcare.**

Paper “reverse engineering”: **Generate a prompt to be input into large language models starting with 'Write a systematic survey or overview about'. [title and abstract of a paper]**

Used survey papers from Semantics Scholar:

1. Computer Science
2. Medicine
3. Chemistry
4. Biology
5. Materials Science
6. Physics
7. Geology
8. Psychology
9. Art
10. History
11. Geography
11. Sociology
12. Business
13. Political Science
14. Economics
15. Philosophy
16. Mathematics
17. Engineering
18. Environmental Science
19. Education
20. Law
21. Linguistics



Data (2): AI-reviewer instructions

- Relevance
- Contribution
- Soundness
- Clarity
- Responsibility

- Overall
- Confidence

Data (3): Generation of semi-synthetic papers

- **Good papers:** **Paraphrased** versions of the human papers

- **Bad papers:**
  - **Clarity:** Rewrite the text with long and windy sentences. Write in poor English.
  - **Contribution:** Act as a person who does not have an analytical mind. Write a bad conclusion.
  - **Responsibility:** Rewrite this text as a political argument, in a polemic style, inappropriate for a scientific paper.

Data (3): Example of “good” and “bad” paraphrased papers

The Role of Artificial Intelligence in Microscopy: A Comprehensive Guide

Deep Learning (DL), a subset of Artificial Intelligence (AI), is paving the way for remarkable advancements in the field of biomedical research, particularly with its potential to transform microscopy.

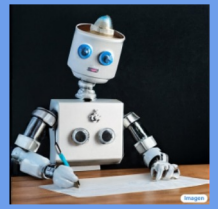
This technology is gradually shifting from computer science professionals to biomedical researchers. [...]

Machine Think for Tiny Look-See: Thing You Need Know

Machine Think with Deep Learn (DL) start new path in life-study research, promise big change for tiny look-see field.

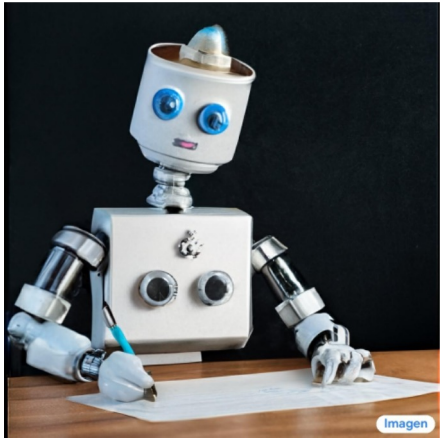
Now, move from computer smart-people hand to life-study research people. [...]

Methods used by the organizers



- Baseline AI-author
- AI referee reviewer
- AI meta-reviewer

Baseline AI author



Ask ChatGPT:

Write a systematic survey or overview about…
Return the result exclusively in this **JSON format**:
[{"heading":"...", "text" : "..."},{"heading":"...", "text" : "..."},...]

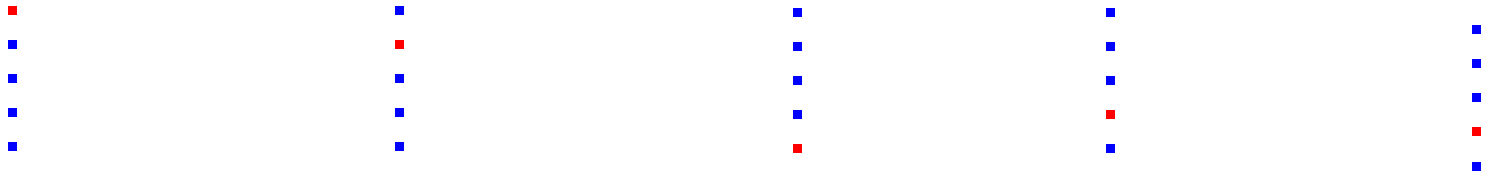
AI referee reviewer

- **Relevance:** (paper title ↔ prompt), (paper abstract ↔ prompt), (citation titles ↔ prompt); based on embedding comparisons
- **Contribution:** (abstract ↔ paper), (conclusion ↔ paper), (citations ↔ reference paper); query language model for first two; last one title comparison.
- **Soundness:** (citations exist), (citations aligned w. paper), (citations diverse); first lookup semantics scholars; last two compare embeddings.
- **Clarity:** (organization), (explanations), (language); Features built from (1) Flesh-Kincaid readability formula and 3 other scores from py-readability-metrics (2) A model predicting Kaggle essay features (cohesion, syntax, vocabulary, phraseology, grammar, conventions.
- **Responsibility:** We used Perspective API to evaluate toxicity. This obviously does not address all ethical issues.

Meta-reviewer (1): contrastive evaluation

- Accuracy of ranking **good** better than **bad** for each criterion

Relevance Contribution Soundness Clarity Responsibility



Score: 1 Score: 0.75 Score: 0 Score: 0.25 Score: 0.25

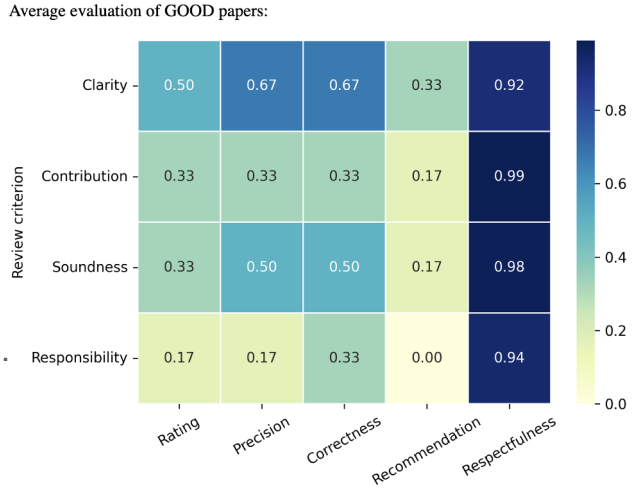
- Review of single AI-generated paper evaluated with overall score.
- Overall scores averaged over multiple generated papers.

Meta-reviewer (2): LLM evaluation

Ask a language model to evaluate the review for each of the review criteria with meta-review criteria

Rating: Is the reviewer score consistent with the text feed-back?
Precision : Is the text feed-back precise (does it point to a specific reason of praise of criticism)?
Correctness: Is the praise or criticism correct and well substantiated?
Recommendation: Does the text feed-back provide detailed and actionable recommendations for improvement?
Respectfulness: Is the language polite and non discriminatory?

Numerical and text evaluation are provided.



Leaderboard score is the average of the contrastive score and the LLM score.

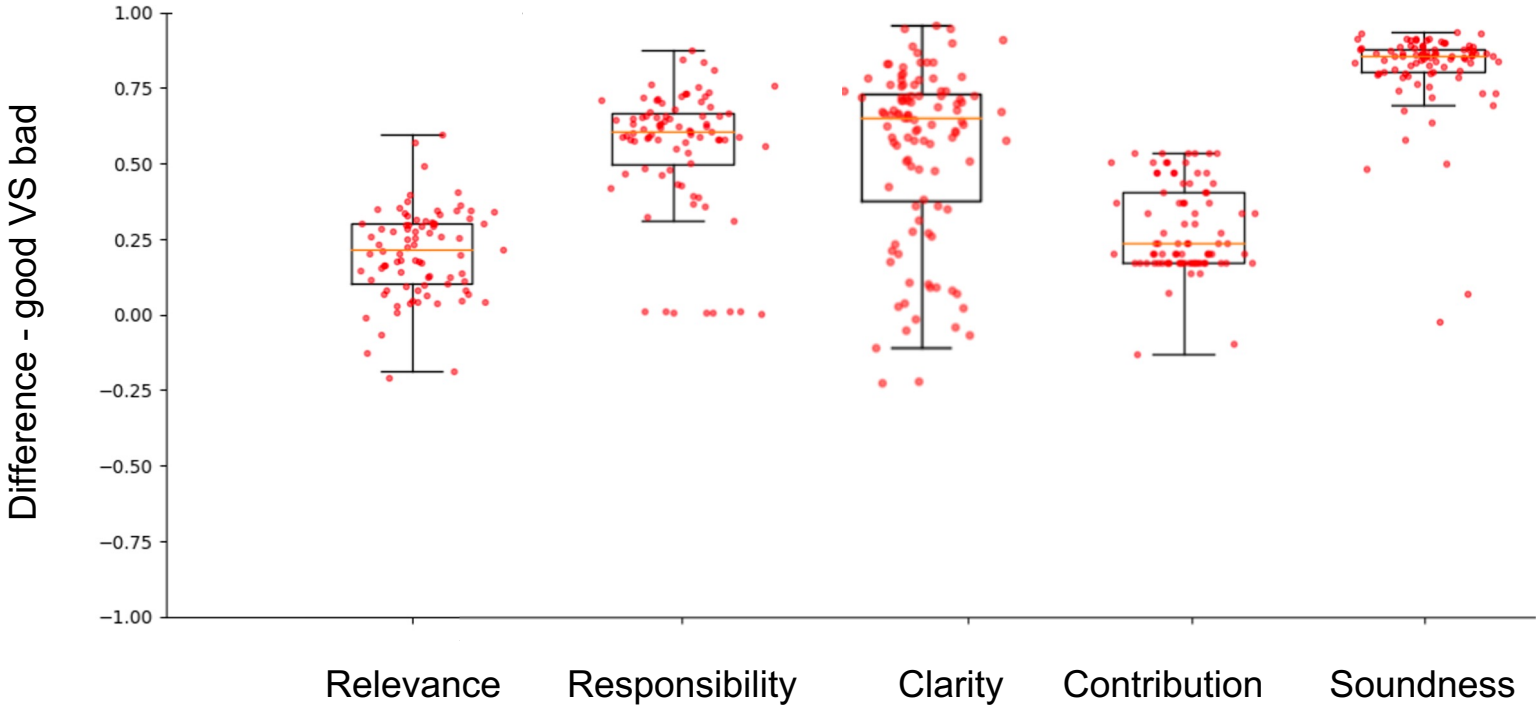
We then average the meta-review score of multiple reviews for 50 papers

Preliminary results

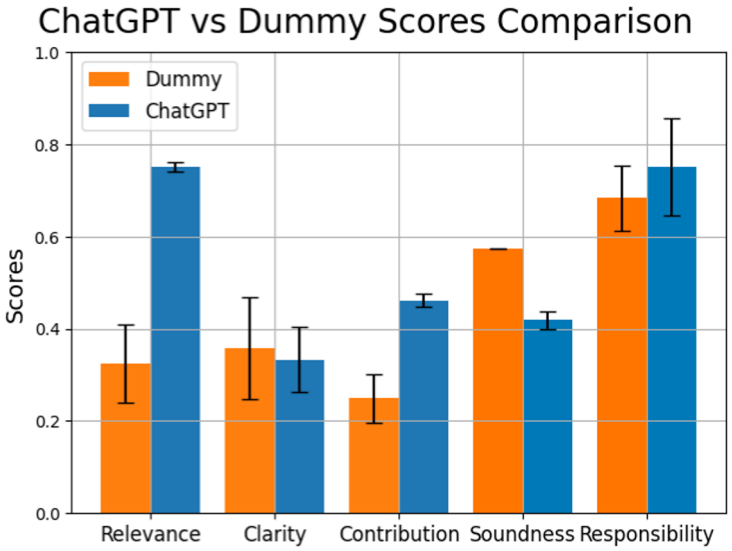
- AI referee reviewer evaluation
  - Good and bad papers
  - AI-generated papers
- AI meta-reviewer sanity check

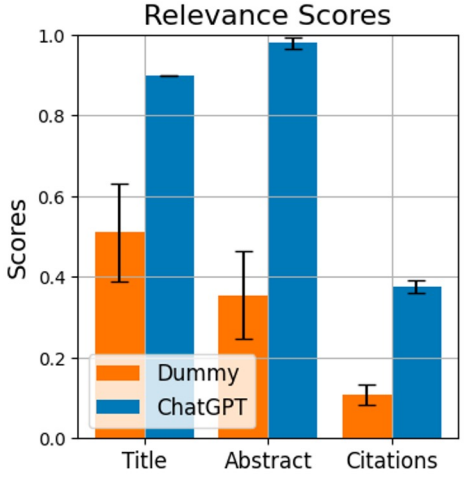
AI referee reviewer evaluation (1): with “ground truth” good and bad papers

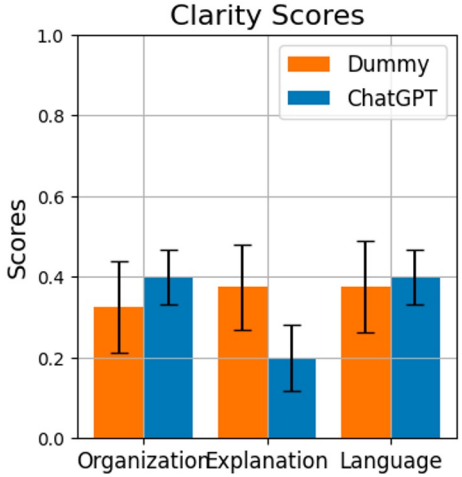
Statistical test to separate “good” and “bad” paraphrased papers for given prompts:

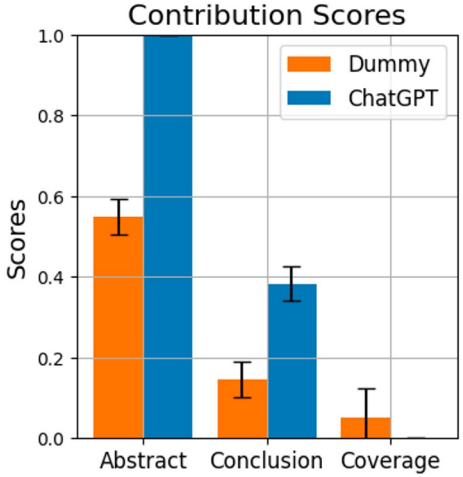


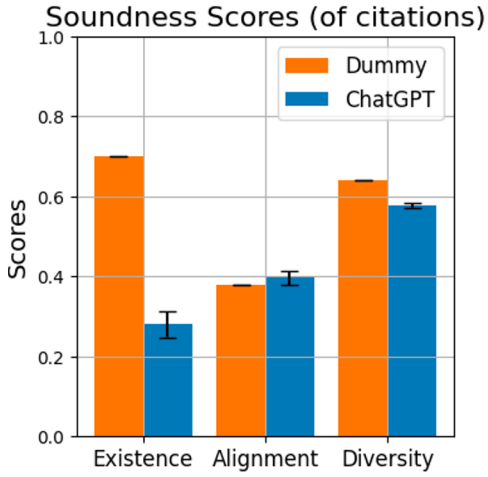
AI referee reviewer evaluation (2): with AI-author generated papers











Meta-reviewer sanity check

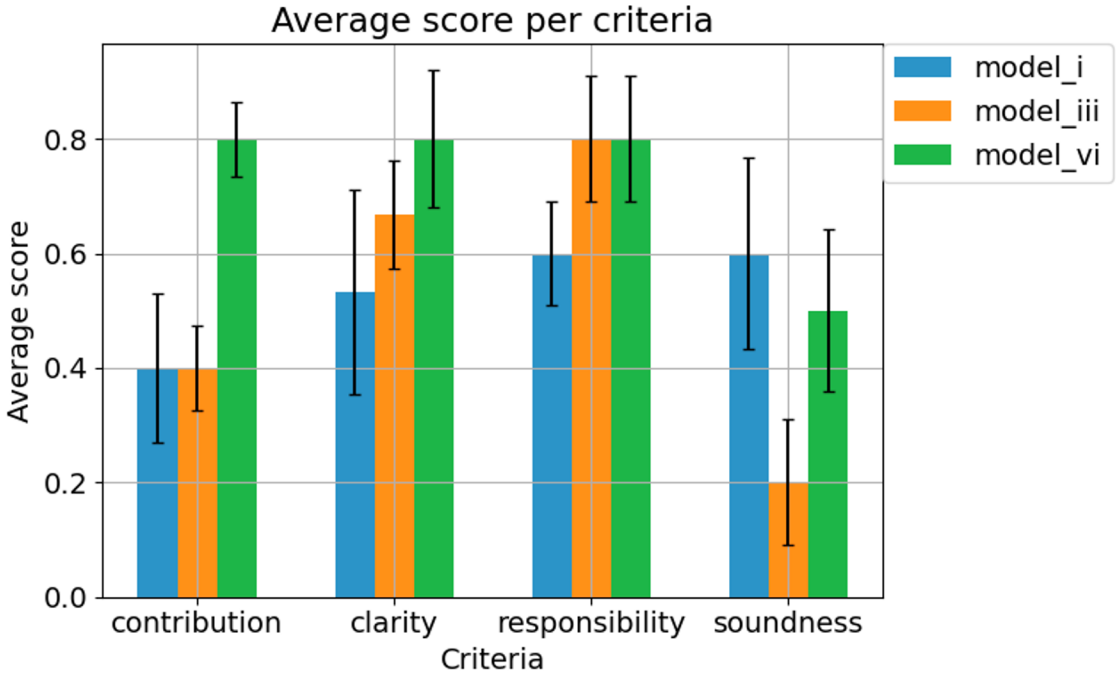
i. **Random-random:** Random review scores, and random supporting text.
ii. **BoW:** A RF classifier using bag-of-words to predict good/bad papers for each criterion.
iii. **LLM plain:** A LLM generates scores and text justifications of the score.
iv. **LLM bogus:** Same as iii, but base the justifications on random scores.
v. **Pairwise comparison:** the LLM reviewer compares the submitted paper with the good and the bad papers and says which one is better w.r.t. the various criteria.
vi. **AI referee reviewer:** Our “hand-crafted” reviewer, with LLM text justification.

Meta-reviewer: Bar graph of sanity check for the contrastive evaluation

i. **Random-random:** 0.48 ± 0.31

iii. **LLM plain:** 0.52 ± 0.21

vi. **AI referee reviewer:** 0.73 ± 0.24



Conclusion & Future work

Status

- Designed end-to-end **competition for peer-reviewed survey papers**
- Implemented preliminary **AI-author, AI-reviewer** AND **AI meta-reviewer** (novel notion)
- Implemented sanity checks for both tracks

Future work

- Improve AI-Author baseline (w.r.t. citations particularly)
- Improve our AI-reviewer (w.r.t. Coverage particularly)
- Further evaluate and improve the meta-reviewer

(in particular evaluate the LLM meta-review)

http://auto-survey.chalearn.org/

Thanks

This is joint work with

Thanh Gia Hieu **KHUONG**

Benedictus Kent **RACHMAT**

Many colleagues contributed ideas and resources:

- To ML datasets and benchmarks
- To Auto-survey, see Credits on http://auto-survey.chalearn.org

http://auto-survey.chalearn.org/

AutoML Conference 2023
Potsdam/Berlin, Germany
September 12th – 15th 2023

Auto-survey tutorial

Thanh Gia Hieu **KHUONG**

Benedictus Kent **RACHMAT**

Advisor: Isabelle **GUYON**

http://auto-survey.chalearn.org/

Search Competitions

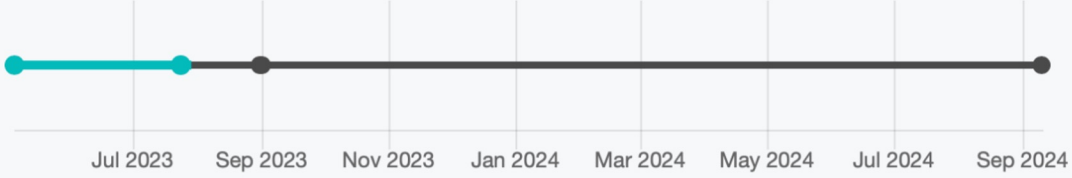
Benchmarks | Resources | Queue Management | Login | Sign-up

AUTO-SURVEY CHALLENGE'23

15

41 SU

ORGANIZED BY: Fnachalearn
CURRENT PHASE ENDS: August 31, 2023, 2:00 AM GMT+2
CURRENT SERVER TIME: July 26, 2023, 2:09 PM GMT+2
Docker image: ktgiahieu/codabench-gpu:paper-challenge



Get Started | Phases | My Submissions | Results | Forum

?

Overview

CHALLENGE'23

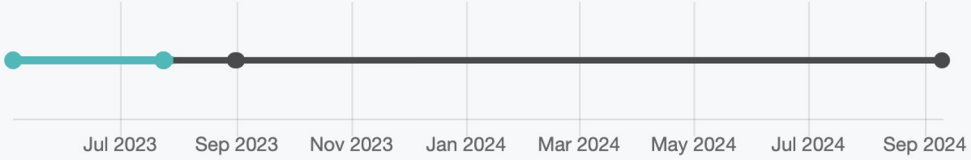
43 SUBMISSIONS

Edit | Participants | Submissions | Dumps | Migrate

ORGANIZED BY: Fnachalearn
CURRENT PHASE ENDS: August 31, 2023 At 2:00 AM GMT+2
CURRENT SERVER TIME: July 26, 2023 At 5:21 PM GMT+2

Docker image: ktgiahieu/codabench-gpu:paper-challenge

Secret url: https://www.codabench.org/competitions/1145/?secret_key=0c9fc8c4-79d3-4dd7-b448-f4e3f05ebd0d



Get Started | Phases | My Submissions | Results | Forum

- Overview
- Data
- Evaluation
- Starting kit
- Prizes

Sample Submission

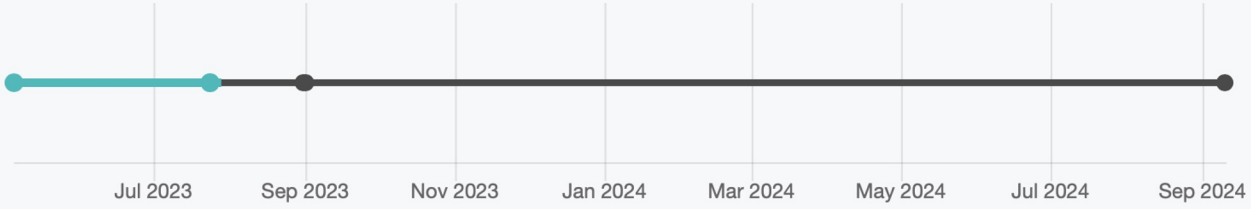
NOTE: The required format of the submission files is described in our starting kit.

📩 "Dummy" Sample Submission

- Generates always the same empty template paper

AUTO-SURVEY CHALLENGE'23

ORGANIZED BY: Fnachalearn
CURRENT PHASE ENDS: August 31, 2023, 2:00 AM GMT+2
CURRENT SERVER TIME: July 26, 2023, 1:52 PM GMT+2
Docker image: ktgiahieu/codabench-gpu:paper-challenge



Get Started | Phases | My Submissions | Results | Forum

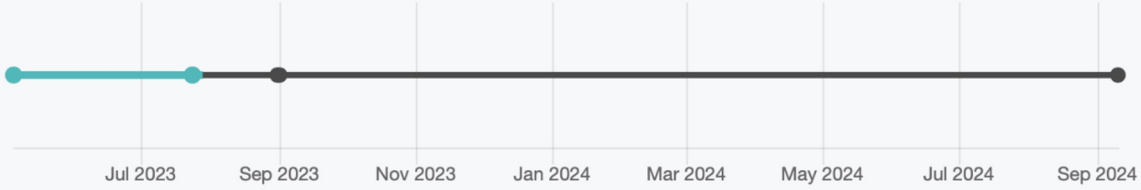
- Overview
- Data
- Evaluation
- Prizes

🤖 Compose and Review Survey Papers

This competition aims at driving progress in the automation and fine-tuning of Large Language Mode
BARD or ***ChatGPT***, with an emphasis on automated prompt engineering.

AUTO-SURVEY CHALLENGE'23

ORGANIZED BY: Fnachalearn
CURRENT PHASE ENDS: August 31, 2023, 2:00 AM GMT+2
CURRENT SERVER TIME: July 26, 2023, 2:11 PM GMT+2
Docker image: ktgiahieu/codabench-gpu:paper-challenge



Get Started | Phases | **My Submissions** | Results | Forum

You have not yet registered for this competition.

To participate in this competition, you must accept its specific terms and conditions.

This competition **requires approval** from the competition organizers. After submitting your registration request, an email will be sent organizers notifying them of your request. Your application will remain pending until they approve or deny it.

☑ I accept the terms and conditions of the competition.

Register

AUTO-SURVEY CHALLENGE'23

17 PARTICIPANTS

41 SUBMISSIONS

ORGANIZED BY: Fnachalearn
CURRENT PHASE ENDS: August 31, 2023, 2:00 AM GMT+2
CURRENT SERVER TIME: July 26, 2023, 2:20 PM GMT+2
Docker image: ktgiahieu/codabench-gpu:paper-challenge

Jul 2023 | Sep 2023 | Nov 2023 | Jan 2024 | Mar 2024 | May 2024 | Jul 2024 | Sep 2024

Get Started | Phases | My Submissions | Results | Forum

Feedback Phase | Development Phase | Final Phase

Submission upload

Metadata or Fact Sheet

Details of your submission: *

DUMMY SAMPLE SUBMISSION

Selected Tasks

- [x] AI-Author
- [x] AI-Reviewer

Submit as:

Yourself

dummy_sample_submission.zip

Get Started | Phases | My Submissions | Results | Forum

Feedback Phase | Development Phase | Final Phase

Submission upload

Metadata or Fact Sheet

Details of your submission: *

DUMMY SAMPLE SUBMISSION

Selected Tasks

- [x] AI-Author
- [x] AI-Reviewer

Submit as: ?

Yourself

Running dummy_sample_submission.zip (ID = 15543)

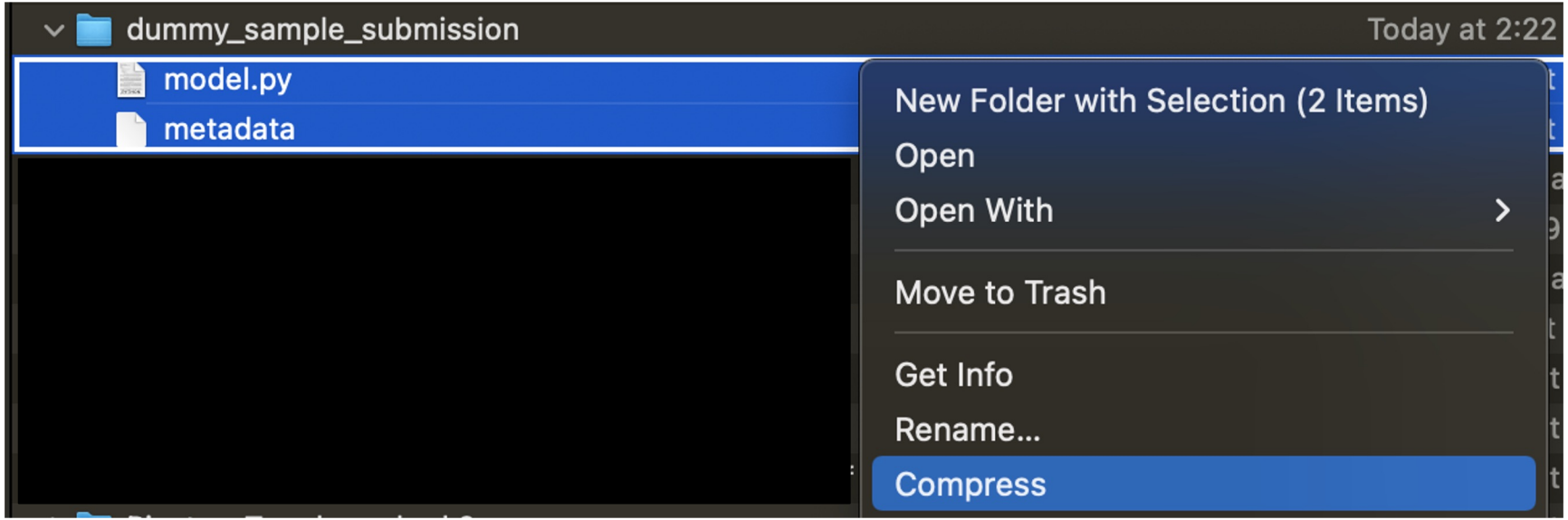
Output

Submission ID: 15544 | Submission ID: 15545

```
Using input_dir: /app/input_data
Using output_dir: /app/output
Using program_dir: /app/program
Using submission_dir: /app/ingested_program
Data name: ai_paper_challenge
Traceback (most recent call last):
File "/app/program/ingestion.py", line 124, in <module>
from model import model
ModuleNotFoundError: No module named 'model'
```

Oh no!!!

Do not zip the top directory



Now it works!

▾ **Running Archive.zip (ID = 15546)**

Output

Submission ID: 15547 | Submission ID: 15548

```
========== Ingestion program version 6 ==========

**************************************************
******** Processing dataset Ai_paper_challenge ********
**************************************************
========= Reading and converting data ==========
###-----------------------------------###
### Checking Data

==================================BUG REPORT==================================
Welcome to bitsandbytes. For bug reports, please run

python -m bitsandbytes
```

Two tracks: AI-AUTHOR-TRACK & AI-REVIEWER-TRACK

Running Arch

Output

Submission ID: 15547 | Submission ID: 15548

```
Using input_dir: /app/input_data
Using output_dir: /app/output
Using program_dir: /app/program
Using submission_dir: /app/ingested_program
Data name: ai_paper_challenge

========== Ingestion program version 6 ==========

**************************************************
******** Processing dataset Ai_paper_challenge ********
**************************************************
========= Reading and converting data ==========
###------------------------------------###
### Checking Data
```

AI Author track

To obtain the scores for the AI Reviewer track, see FAQ.

Below are the scores of your submission, obtained after AI-referee-reviewer review **5** papers.

Detailed reviews:

1. Prompt 1
2. Prompt 2
3. Prompt 3
4. Prompt 4

Running Archive.zip (ID = 15546)

Output

Submission ID: 15547 | Submission ID: 15548

```
Using input_dir: /app/input_data
Using output_dir: /app/output
Using program_dir: /app/program
Using submission_dir: /app/ingested_program
Data name: ai_paper_challenge

========== Ingestion program version 6 ==========

**************************************************
******** Processing dataset Ai_paper_challenge ********
**************************************************
========= Reading and converting data ==========
###------------------------------------###
### Checking Data
```

AI-Reviewer track

Below are the scores of your submission, obtained by our meta-reviewer after reviewing **90** papers

Average meta-reviewer evaluation of your reviewer's NUMERICAL scores:

The scores below represent for each criterion the accuracy of rating good papers better than bad papers

Ready for a “real” submission? Use ChatGPT!

Get Started | Phases | My Submissions | Results | Forum

- Overview
- Data
- Evaluation
- Starting kit
- Prizes
- Terms and Conditions
- FAQ
- Files

Sample Submission

NOTE: The required format of the submission files is described in our starting kit.

"Dummy" Sample Submission

- Generates always the same empty template paper
- Return 0 score with no comment for every review

If you happen to unzip this sample submission, beware that, to submit it on Codabench, you need to re-zip it WITHOUT the top directory.

ChatGPT Sample Submission

- Generates papers using ChatGPT
- Reviews papers using ChatGPT

Need your OpenAI API key

You will need to include an OpenAI API key in your submission, please follow the instruction here to obtain one. Although you can run the submission on your local computer (recommended if you make changes and want to try out new ideas), all you need to do to run this submission on Codabench is to replace the OpenAI API key in sample_submission_chatgpt_api_key.json and re-zip the contents of the sample submission (WITHOUT the top directory).

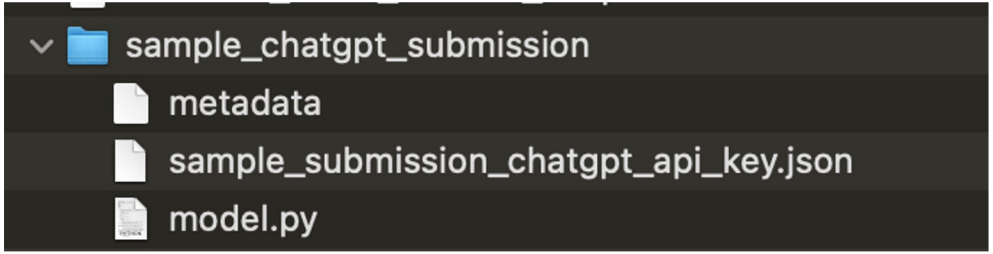
Starting Kit

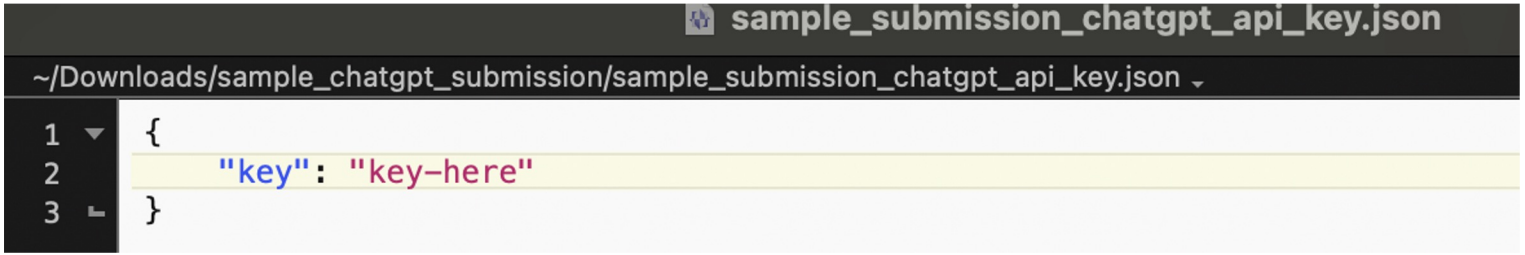
Open in Colab

You can also take advantage of our starting kit for a quick initiation that runs in Google Colab. This kit includes sample_data/ and guidelines to help you understand the data, construct your API (using

Get your OpenAI API key

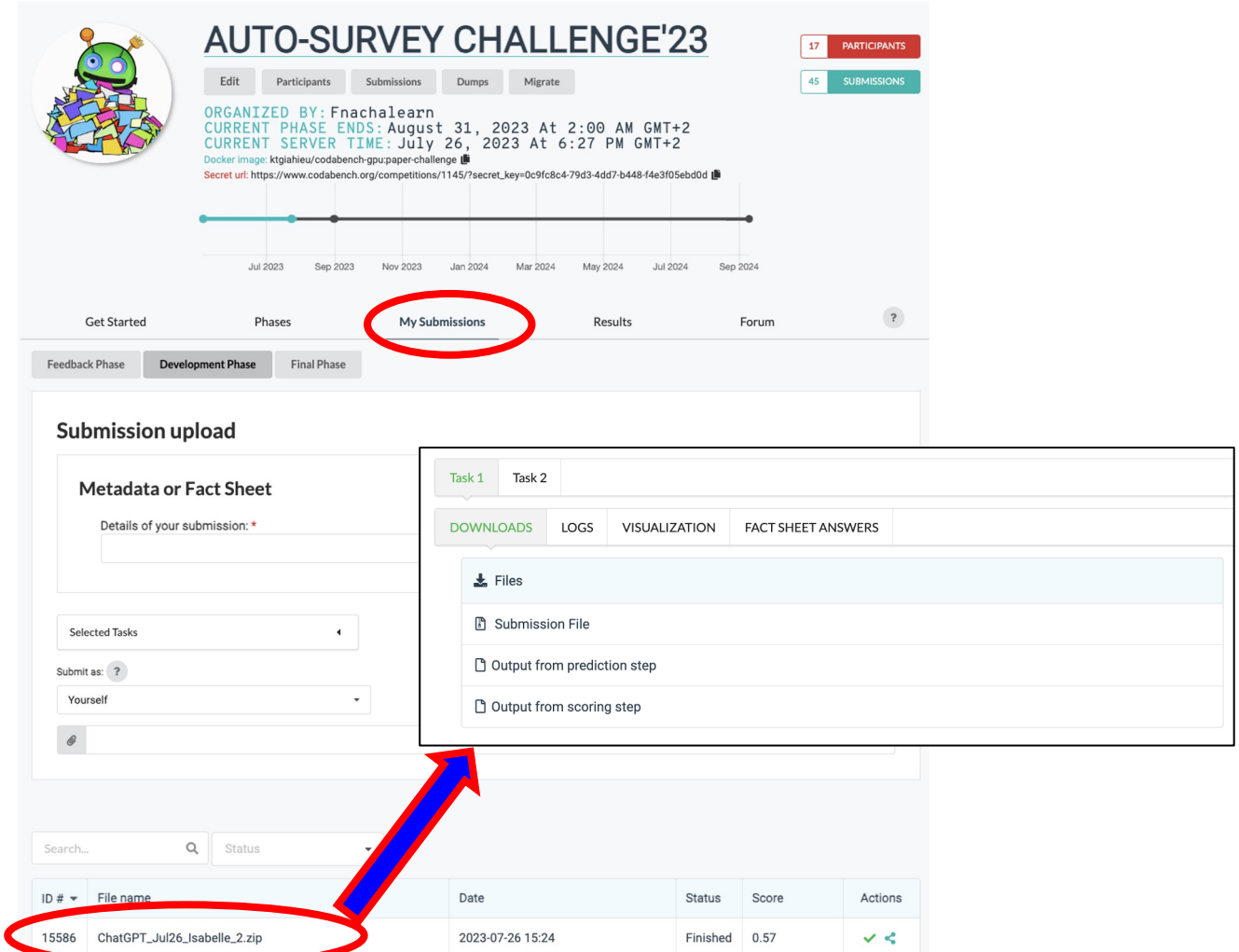
https://github.com/kentrachmat/public_ai_paper_challenge_codabench/blob/master/how_to_get_openai_api.md





Re-zip without the top directory and submit to Codabench.

Ta dah!


AUTO-SURVEY CHALLENGE'23
17 PARTICIPANTS
45 SUBMISSIONS
Edit
Participants
Submissions
Dumps
Migrate
ORGANIZED BY: Fnachalearn
CURRENT PHASE ENDS: August 31, 2023 At 2:00 AM GMT+2
CURRENT SERVER TIME: July 26, 2023 At 6:27 PM GMT+2
Docker image: ktgiahieu/codabench-gpu:paper-challenge
Secret url: https://www.codabench.org/competitions/1145/?secret_key=0c9fc8c4-79d3-4dd7-b448-f4e3f05ebd0d
Jul 2023
Sep 2023
Nov 2023
Jan 2024
Mar 2024
May 2024
Jul 2024
Sep 2024
Get Started
Phases
My Submissions
Results
Forum
Feedback Phase
Development Phase
Final Phase
Submission upload
Metadata or Fact Sheet
Details of your submission: *
Selected Tasks
Submit as:
Yourself
Task 1
Task 2
DOWNLOADS
LOGS
VISUALIZATION
FACT SHEET ANSWERS
Files
Submission File
Output from prediction step
Output from scoring step
Search...
Status
ID #
File name
Date
Status
Score
Actions
15586
ChatGPT_Jul26_Isabelle_2.zip
2023-07-26 15:24
Finished
0.57

Leaderboard

Get Started | Phases | My Submissions | Results | Forum

Feedback Phase | Development Phase | Final Phase

Filter Leaderboard by Columns

Results

Task:				AI-Author		AI-Reviewer	
#	Participant	Entries	Date of last entry	score	Detailed Results	score	Detailed Results
1	guyon	2	2023-07-26	n/a		0.57	

Detailed Results

AI-Reviewer track

Below are the scores of your submission, obtained by our meta-reviewer after reviewing **90** papers

Average meta-reviewer evaluation of your reviewer's NUMERICAL scores:

The scores below represent for each criterion the accuracy of rating good papers better than bad papers

responsibility: 0.80
contribution: 0.29
clarity: 0.67
soundness: 0.10

Overall score: 0.46

Red means score below 0.5.

Box plot of score diffence between good and bad papers:

Starting kit

Get Started | Phases | My Submissions | Results | Forum | ?

- Overview
- Data
- Evaluation
- **Starting kit**
- Prizes
- Terms and Conditions
- FAQ
- Files

Sample Submission

NOTE: The required format of the submission files is described in our starting kit.

📩 "Dummy" Sample Submission

- Generates always the same empty template paper
- Return 0 score with no comment for every review

If you happen to unzip this sample submission, beware that, to submit it on Codabench, you need to re-zip it WITHOUT the top directory.

📩 ChatGPT Sample Submission

- Generates papers using ChatGPT
- Reviews papers using ChatGPT

You will need to include an OpenAI API key in your submission, please follow the instruction here to obtain one. Although you can run the submission on your local computer (recommended if you make changes and want to try out new ideas), all you need to do to run this submission on Codabench is to replace the OpenAI API key in sample_submission_chatgpt_api_key.json and re-zip the contents of the sample submission (WITHOUT the top directory).

🎁 Starting Kit

Open in Colab

Survey Paper Challenge 2023 - Starting Kit

File Edit View Insert Runtime Tools Help Cannot save changes

+ Code + Text Copy to Drive

Starting Kit: AUTO-SURVEY CHALLENGE'23

This starting kit will guide you step by step and will walk you through the data statistics and examples. This will give you a clear idea of what this challenge is about and how you can proceed further to solve the challenge.

- Prerequisite
- Technical Details
- Disclaimer
- References and Credits

Competition Site | Landing Page Site | Contact us

Step 0: Input your OpenAI API key

To run this notebook, you must get an OpenAI API key by following this instruction.

```
import getpass
from IPython.display import clear_output

clear_output()
# Please enter your API key
new_api_key = ''
while len(new_api_key) == 0:
    new_api_key = getpass.getpass("Please input your API key: ")
    clear_output()
```

Step 1: Import required packages

```
!git clone -q https://github.com/kentrachmat/public_ai_paper_challenge_codabench.git
%cd public_ai_paper_challenge_codabench
!pip install -q openai tiktoken json5
!python -m nltk.downloader punkt
```

```
/content/public_ai_paper_challenge_codabench/public_ai_paper_challenge_codabench/public_ai_paper_challenge_codabench
/usr/lib/python3.10/runpy.py:126: RuntimeWarning: 'nltk.downloader' found in sys.modules after import of package 'nltk',
  warn(RuntimeWarning(msg))
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
```

0s completed at 5:29 PM