Are LLMs Any Good for Scientific Leaderboard Extraction?

Moritz Staudinger*, *Alaa El-Ebshihy**, *Annisa Maulida Ningtyas**, *Florina Piroi*, *Allan Hanbury*

CLEF 2024 GRENOBLE

10th of September, 2024, Grenoble, France

TU WIEN Informatics

DATA SCIENCE INSIGHTS FOR SOCIETY

RSA FG Research Studios Austria Forschungsgesellschaft

What are Scientific Leaderboards?

- Compare Scientific Results
  - Task
  - Dataset
  - Metric

- Find best models for given task

- Currently manually curated

Example: “Efficient Adaptive Ensembling for Image Classification”

Task: Image Classification (from title)

Dataset: CIFAR-10, CIFAR-100, …

Metric: Accuracy, Improvement

Score: 99.5%, 99.612%, 0.112%, …

→ Manually extract these TDMS combinations for leaderboards

Dataset	SOTA accuracy	Our accuracy	Improvement
CIFAR-10 [30]	99.500%	99.612%	0.112%
CIFAR-100 [31]	96.080%	96.808%	0.728%
Cars [32]	96.320%	96.868%	0.548%
Food-101 [31]	96.180%	96.879%	0.699%
Flower102 [33]	99.720%	99.847%	0.127%
CINIC-10 [34]	94.300%	95.064%	0.764%
Pets [31]	97.100%	98.220%	1.120%

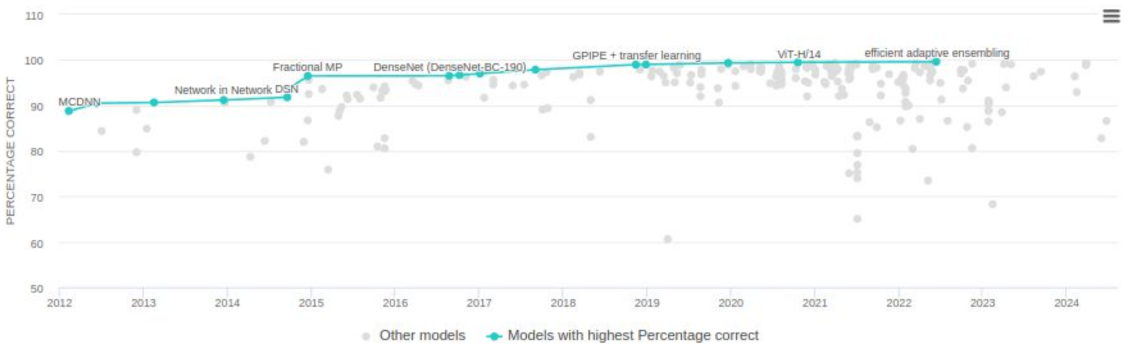
In order to stress our method, we also provide a different combination of weak classifiers: specifically, we show the results of an ensemble of five weak models. For demonstration purposes we report the results obtained only for the CIFAR-100 and CIFAR-10 datasets. In the case of CIFAR-100, while the ensemble using 2 weak models obtained an accuracy of 96.808%, the new one obtained an accuracy of 84.930%. This result was expected, since each weak model had to be trained on a third of the images of the previous case according to the data splitting procedure described in Section 4.6 in order to avoid the use of the same images. In the case of CIFAR-10, while the ensemble using 2 weak models obtained an accuracy of 99.612%, the new one obtained an accuracy of 96.640%

Example: "Efficient Adaptive Ensembling for Image Classification"

Image Classification on CIFAR-10

Leaderboard   Dataset

View [Percentage correct] by [Date] for [All models]



Filter: Transformer | ResNet | EfficientNet | Xformer | SNN | DCN | TM | Memory-Centric | CLIP Pre-trained | ResNeXt | ResNet-18 | untagged

Edit Leaderboard

Rank	Model	Percentage correct	PARAMS	Extra Training Data	Paper	Code	Result	Year	Tags
1	**efficient adaptive ensembling**	99.612		✓	Efficient Adaptive Ensembling for Image Classification			2022	
2	**ViT-H/14**	99.5	632M	✓	An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale			2020	Transformer
3	**DINOv2** (ViT-g/14, frozen model, linear eval)	99.5	1100M	✓	DINOv2: Learning Robust Visual Features without Supervision			2023	
4	**µ2Net** (ViT-L/16)	99.49		✓	An Evolutionary Approach to Dynamic Introduction of Tasks in Large-scale Multitask Learning Systems			2022	

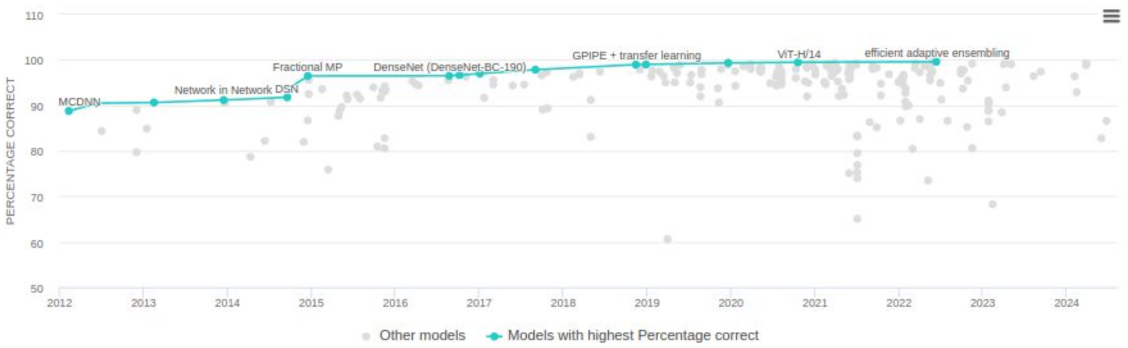
https://paperswithcode.com/sota/image-classification-on-cifar-10

Example: “Efficient Adaptive Ensembling for Image Classification”

Image Classification on CIFAR-10

Leaderboard Dataset

View: Percentage correct by Date for All models



Filter: Transformer ResNet EfficientNet Xformer SNN DCN TM Memory-Centric CLIP Pre-trained ResNeXt ResNet-18 untagged

Edit Leaderboard

Rank	Model	Percentage correct	PARAMS	Extra Training Data	Paper	Code	Result	Year	Tags
1	**efficient adaptive ensembling**	99.612		✓	Efficient Adaptive Ensembling for Image Classification			2022	
2	**ViT-H/14**	99.5	632M	✓	An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale			2020	Transformer
3	**DINOv2** (ViT-g/14, frozen model, linear eval)	99.5	1100M	✓	DINOv2: Learning Robust Visual Features without Supervision			2023	
4	**µ2Net** (ViT-L/16)	99.49		✓	An Evolutionary Approach to Dynamic Introduction of Tasks in Large-scale Multitask Learning Systems			2022	

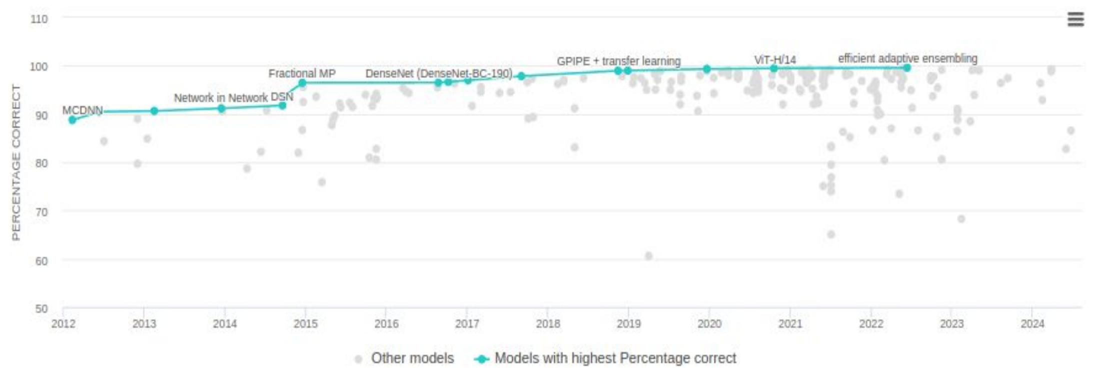
https://paperswithcode.com/sota/image-classification-on-cifar-10

Image Classification

Image Classification on CIFAR-10

Leaderboard Dataset

View Percentage correct by Date for All models



Filter: Transformer ResNet EfficientNet Xformer SNN DCN TM Memory-Centric CLIP Pre-trained ResNeXt ResNet-18 untagged

Edit Leaderboard

Rank	Model	Percentage correct	PARAMS	Extra Training Data	Paper	Code	Result	Year	Tags
1	**efficient adaptive ensembling**	99.612		✓	Efficient Adaptive Ensembling for Image Classification			2022	
2	**ViT-H/14**	99.5	632M	✓	An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale			2020	Transformer
3	**DINOv2** (ViT-g/14, frozen model, linear eval)	99.5	1100M	✓	DINOv2: Learning Robust Visual Features without Supervision			2023	
4	**µ2Net** (ViT-L/16)	99.49		✓	An Evolutionary Approach to Dynamic Introduction of Tasks in Large-scale Multitask Learning Systems			2022	

https://paperswithcode.com/sota/image-classification-on-cifar-10

Dataset	SOTA accuracy	Our accuracy	Improvement
CIFAR-10 [30]	99.500%	99.612%	0.112%
CIFAR-100 [31]	96.080%	96.808%	0.728%
Cars [32]	96.320%	96.868%	0.548%
Food-101 [31]	96.180%	96.879%	0.699%
Flower102 [33]	99.720%	99.847%	0.127%
CINIC-10 [34]	94.300%	95.064%	0.764%
Pets [31]	97.100%	98.220%	1.120%

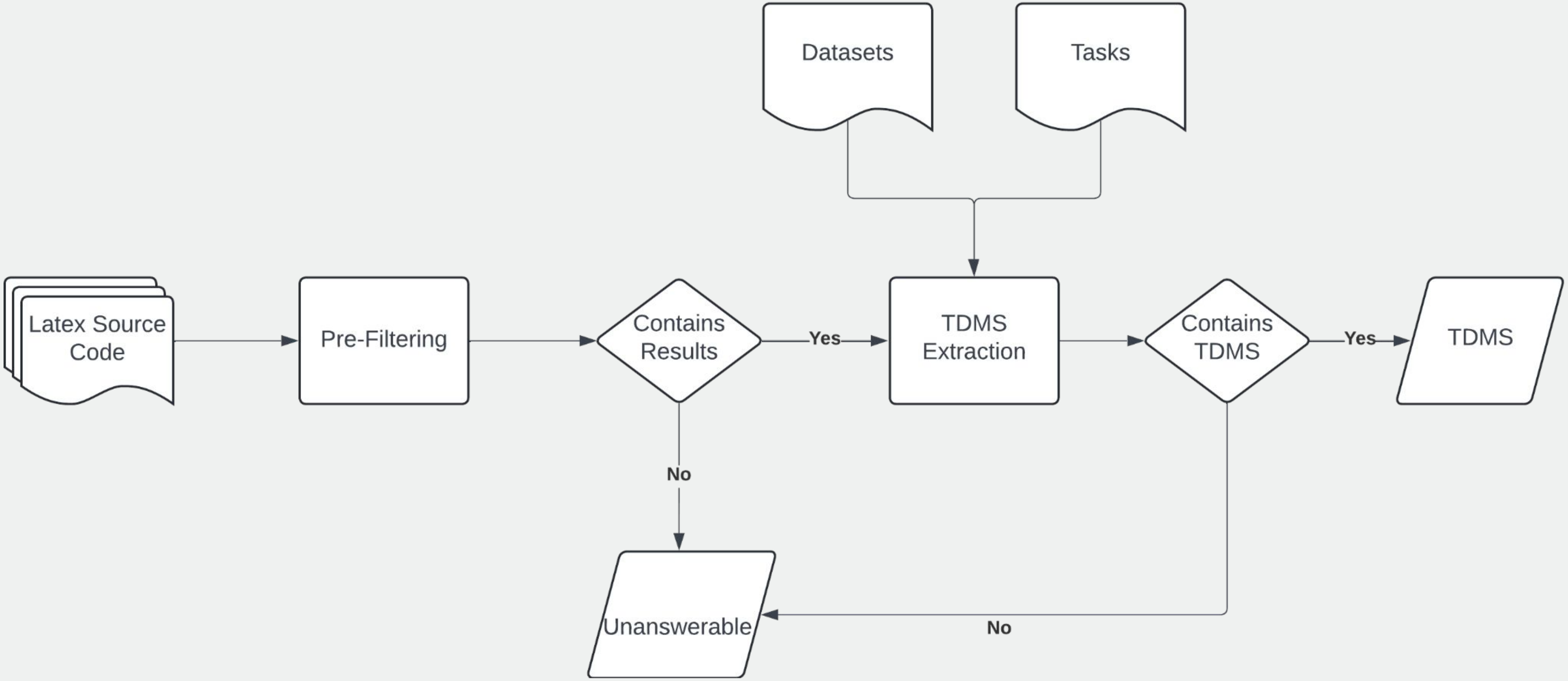
In order to stress our method, we also provide a different combination of weak classifiers: specifically, we show the results of an ensemble of five weak models. For demonstration purposes we report the results obtained only for the CIFAR-100 and CIFAR-10 datasets. In the case of CIFAR-100, while the ensemble using 2 weak models obtained an accuracy of 96.808%, the new one obtained an accuracy of 84.930%. This result was expected, since each weak model had to be trained on a third of the images of the previous case according to the data splitting procedure described in Section 4.6 in order to avoid the use of the same images. In the case of CIFAR-10, while the ensemble using 2 weak models obtained an accuracy of 99.612%, the new one obtained an accuracy of 96.640%

https://arxiv.org/pdf/2206.07394v3

Task

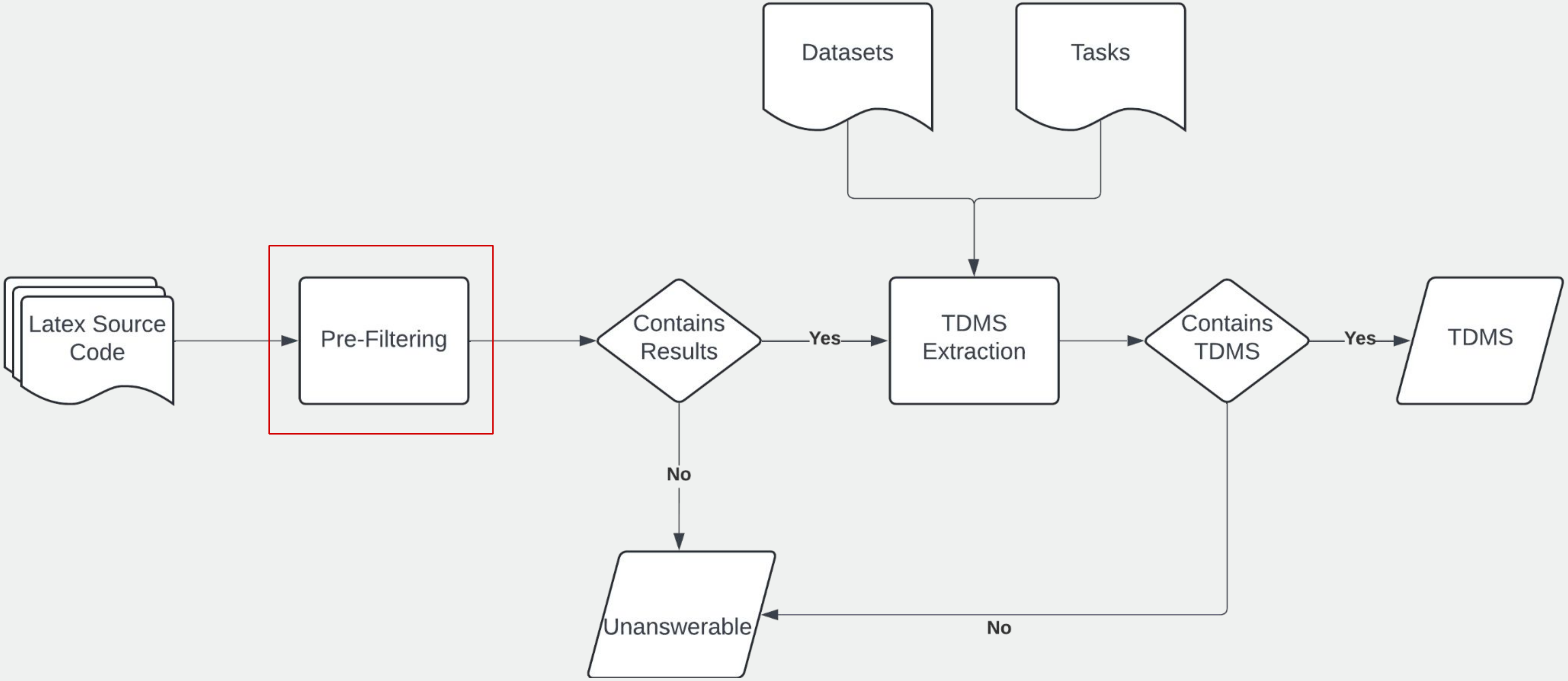
Develop a machine learning model that can distinguish whether a scholarly article provided as input to the model reports a TDMS or not. And for articles reporting TDMSs, extract all the relevant ones.

Process



Datasets
Tasks
Latex Source Code
Pre-Filtering
Contains Results
Yes
TDMS Extraction
Contains TDMS
Yes
TDMS
No
Unanswerable
No

Process


Datasets
Tasks
Latex Source Code
Pre-Filtering
Contains Results
Yes
TDMS Extraction
Contains TDMS
Yes
TDMS
No
Unanswerable
No

Rule Based Pre-Filtering

- Rule-based binary classifier → LaTeX source code structures
- Avoid complex models for simple tasks
- Recall-Oriented

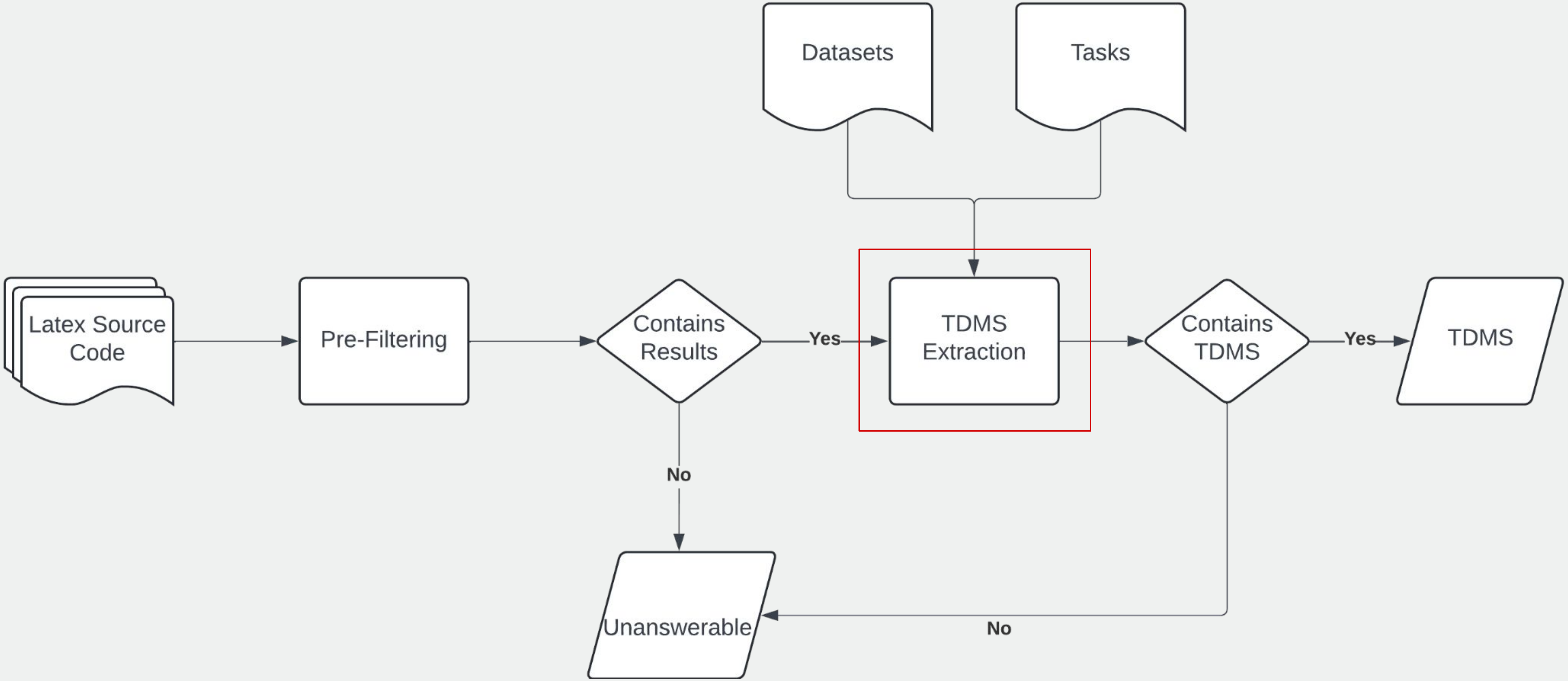
Method
Result Section Exists
Result Section Exists with add. terms
Result Table Exists

Rule Based Pre-Filtering

- Rule-based binary classifier → LaTeX source code structures
- Avoid complex models for simple tasks
- Recall-Oriented

Method	Precision	Recall	Accuracy
Result Section Exists	0.685	0.96	0.76
Result Section Exists with add. terms	0.67	0.98	0.75
Result Table Exists	0.85	0.80	0.83

Process


Datasets
Tasks
Latex Source Code
Pre-Filtering
Contains Results
Yes
No
TDMS Extraction
Contains TDMS
Yes
TDMS
Unanswerable
No

TDMS Extraction

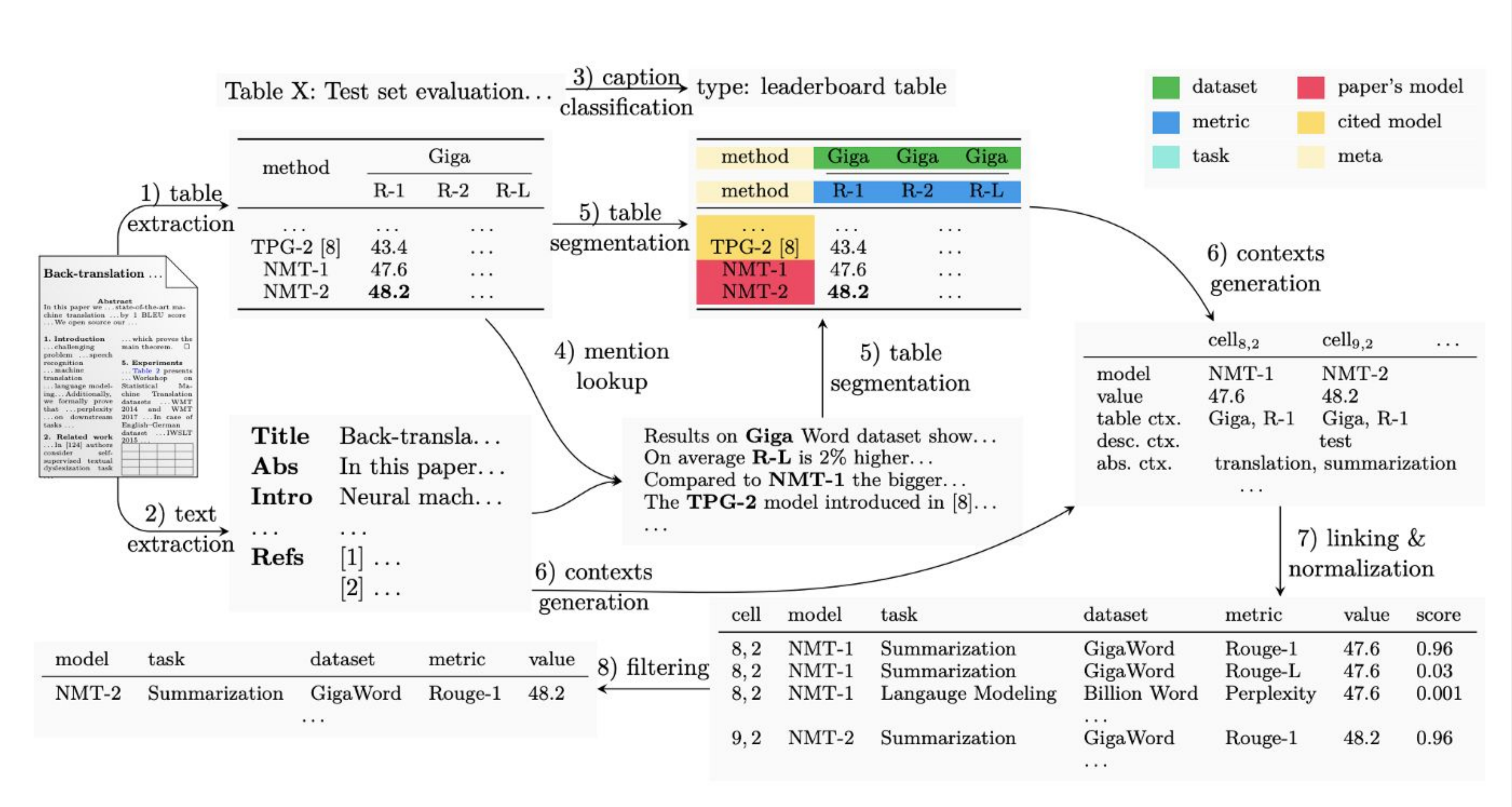
Type	Id	Name	Filtered	zero- or few-shot	fulltext or az	PwC information
Baseline	1	AxCell	✓	–	fulltext	✗

TDMS Extraction

Type	Id	Name	Filtered	zero- or few-shot	fulltext or az	PwC information
Baseline	1	AxCell	✓	–	fulltext	✗
LLMs	2	GPT35-zero	✗	zero	fulltext	✗
	3	GPT35-fil-zero	✓	zero	fulltext	✗
	4	GPT35-few	✗	few	fulltext	✗
	5	GPT35-fil-few	✓	few	fulltext	✗
	6	GPT35-info-few	✗	few	fulltext	✓
	7	GPT35-az-few	✗	few	az	✗
	8	GPT35-az-info-few	✗	few	az	✓
	9	Mistral-fil-zero	✓	zero	fulltext	✗
	10	Mistral-fil-info-zero	✓	zero	fulltext	✓

TDMS Extraction

- ML Pipeline
  - Table Extraction
  - Text Extraction
- Result Merging
- Uses Arxiv as source → overlap in the collections



TDMS Extraction

Type	Id	Name	Filtered	zero- or few-shot	fulltext or az	PwC information
LLMs	2	GPT35-zero	✗	zero	fulltext	✗
	3	GPT35-fil-zero	✓	zero	fulltext	✗
	4	GPT35-few	✗	few	fulltext	✗
	5	GPT35-fil-few	✓	few	fulltext	✗
	6	GPT35-info-few	✗	few	fulltext	✓
	7	GPT35-az-few	✗	few	az	✗
	8	GPT35-az-info-few	✗	few	az	✓
	9	Mistral-fil-zero	✓	zero	fulltext	✗
	10	Mistral-fil-info-zero	✓	zero	fulltext	✓

zero shot

few shots

TDMS Extraction

Type	Id	Name	Filtered	zero- or few-shot	fulltext or az	PwC information
LLMs	2	GPT35-zero	✗	zero	fulltext	✗
	3	GPT35-fil-zero	✓	zero	fulltext	✗
	4	GPT35-few	✗	few	fulltext	✗
	5	GPT35-fil-few	✓	few	fulltext	✗
	6	GPT35-info-few	✗	few	fulltext	✓
	7	GPT35-az-few	✗	few	az	✗
	8	GPT35-az-info-few	✗	few	az	✓
	9	Mistral-fil-zero	✓	zero	fulltext	✗
	10	Mistral-fil-info-zero	✓	zero	fulltext	✓

unanswerable
filtered

TDMS Extraction

Type	Id	Name	Filtered	zero- or few-shot	fulltext or az	PwC information
LLMs	2	GPT35-zero	✗	zero	fulltext	✗
	3	GPT35-fil-zero	✓	zero	fulltext	✗
	4	GPT35-few	✗	few	fulltext	✗
	5	GPT35-fil-few	✓	few	fulltext	✗
	6	GPT35-info-few	✗	few	fulltext	✓
	7	GPT35-az-few	✗	few	az	✗
	8	GPT35-az-info-few	✗	few	az	✓
	9	Mistral-fil-zero	✓	zero	fulltext	✗
	10	Mistral-fil-info-zero	✓	zero	fulltext	✓

results and experiments sections
(Argumentative Zoning)

TDMS Extraction

Type	Id	Name	Filtered	zero- or few-shot	fulltext or az	PwC information
LLMs	2	GPT35-zero	✗	zero	fulltext	✗
	3	GPT35-fil-zero	✓	zero	fulltext	✗
	4	GPT35-few	✗	few	fulltext	✗
	5	GPT35-fil-few	✓	few	fulltext	✗
	6	GPT35-info-few	✗	few	fulltext	✓
	7	GPT35-az-few	✗	few	az	✗
	8	GPT35-az-info-few	✗	few	az	✓
	9	Mistral-fil-zero	✓	zero	fulltext	✗
	10	Mistral-fil-info-zero	✓	zero	fulltext	✓

PwC
additional information

Results

The Accuracy and Summary results of our Phase 2 submissions

Model	Accuracy	Summary			
		Rouge 1	Rouge 2	Rouge L	Rouge Lsum
AxCell	**83.4**	**75.25**	**4.56**	**74.85**	**73.7**
GPT35-fil-zero	67.05	66.61	0.11	66.54	66.44
GPT35-few	**85.93**	**73.72**	**6.07**	**72.72**	**72.57**
GPT35-fil-few	69.07	68.81	0.14	68.76	68.66
GPT35-info-few	72.75	59.22	2.48	59.06	58.99
GPT35-az-few	**79.09**	**71.07**	**3.56**	**70.82**	**70.62**
GPT35-az-info-few	**75.41**	**71.59**	1.71	**71.46**	**71.35**
Mistral-fil-zero	**75.79**	**68.92**	2.18	**67.51**	**66.48**
Mistral-fil-info-zero	71.23	56.63	3.7	55.14	53.09

Drop because misleading information

Results

The overall results of all TDMS for our Phase 2 submissions

Model	Exact			Inexact		
	P	R	F1	P	R	F1
AxCell	**36.36**	**6.21**	**10.6**	**40.85**	**6.97**	**11.9**
GPT35-fil-zero	2.63	0.15	0.29	3.08	0.18	0.34
GPT35-few	**12.82**	**9.89**	**11.16**	**16.74**	**12.81**	**14.52**
GPT35-fil-few	4.04	0.15	0.3	4.04	0.15	0.3
GPT35-info-few	4.46	2.24	2.99	6.82	3.43	4.56
GPT35-az-few	**11.78**	**3.77**	**5.71**	**18.16**	**5.7**	**8.68**
GPT35-az-info-few	**13.45**	**1.38**	2.5	**23.12**	**2.33**	4.23
Mistral-fil-zero	**8.05**	**4.42**	**5.71**	**10.74**	**5.89**	**7.61**
Mistral-fil-info-zero	**11.64**	**8.37**	**9.74**	**14.62**	**10.5**	**12.22**

Increase because valuable information

Manual Analysis

Sample output of TDMS extracted from Arxiv ID 2012.02818v2

Data source	Task-Dataset-Metric-Score
Ground Truth	Image Classification - CIFAR-10 - Percentage correct - 95.02 Image Classification - CIFAR-100 - Percentage correct - 76.85
AxCell	Image Classification - CIFAR-10 - Percentage error - 4.98 Image Classification - CIFAR-100 - Percentage error - 20.7 Semantic Segmentation - KITTI Semantic Segmentation - Mean IoU (class) - 89.08
GPT35-zero	Image classification - CIFAR-10, CIFAR-100 - Accuracy - 96.74% Image classification - CIFAR-10, CIFAR-100 - AUC - 0.9803 Stability analysis of Bayesian Neural Networks - CIFAR-10 - Epoch Divergence - None OOD detection - CIFAR-10, CIFAR-100 - ECE - 0.0520 Semantic segmentation - StreetHazards, BDD-Anomaly - mIoU - 56.12%

Manual Analysis

Sample output of TDMS extracted from Arxiv ID 2012.02818v2

Data source	Task-Dataset-Metric-Score
Ground Truth	Image Classification - CIFAR-10 - Percentage correct - 95.02 Image Classification - CIFAR-100 - Percentage correct - 76.85
AxCell	Image Classification - CIFAR-10 - Percentage error - 4.98 Image Classification - CIFAR-100 - Percentage error - 20.7 Semantic Segmentation - KITTI Semantic Segmentation - Mean IoU (class) - 89.08
GPT35-zero	Image classification - CIFAR-10, CIFAR-100 - Accuracy - 96.74% Image classification - CIFAR-10, CIFAR-100 - AUC - 0.9803 Stability analysis of Bayesian Neural Networks - CIFAR-10 - Epoch Divergence - None OOD detection - CIFAR-10, CIFAR-100 - ECE - 0.0520 Semantic segmentation - StreetHazards, BDD-Anomaly - mIoU - 56.12%

Conclusion

- No reliable results yet
  → Repeat runs

- Test data needs improvement
  - Standardized naming
  - All results should be included

- External knowledge beneficial for TDMS extraction
  → Cause hallucinations in classification task

- Only <60% of TDMS values are in preprints
  → How does the coverage change for the actual papers?

Any Questions?