SimpleText Task 2

Identify and Explain Difficult Concepts

Giorgio Maria Di Nunzio, Federica Vezzani, Vanessa Bonato, Hosein Azarbonyad,
Jaap Kamps and Liana Ermakova

9 September 2024, CLEF 2024, Grenoble, France

SimpleText Task 2

Identify and Explain Difficult Concepts

- **Understanding terminology** is crucial for comprehending scientific information.
    - Comprehension of the term implies grasping the concept it represents without the need for an explicit definition.
- **Writing clear definitions** of scientific terms makes complex concepts more understandable.
    - Providing accurate definitions and background knowledge can reduce the risk of misinterpreting scientific information

Goal of Task 2

- The goal of this task is to **identify key concepts** that need to be **contextualized with a definition**, example or use case, and provide useful and understandable explanations for them.

- There are three subtasks:

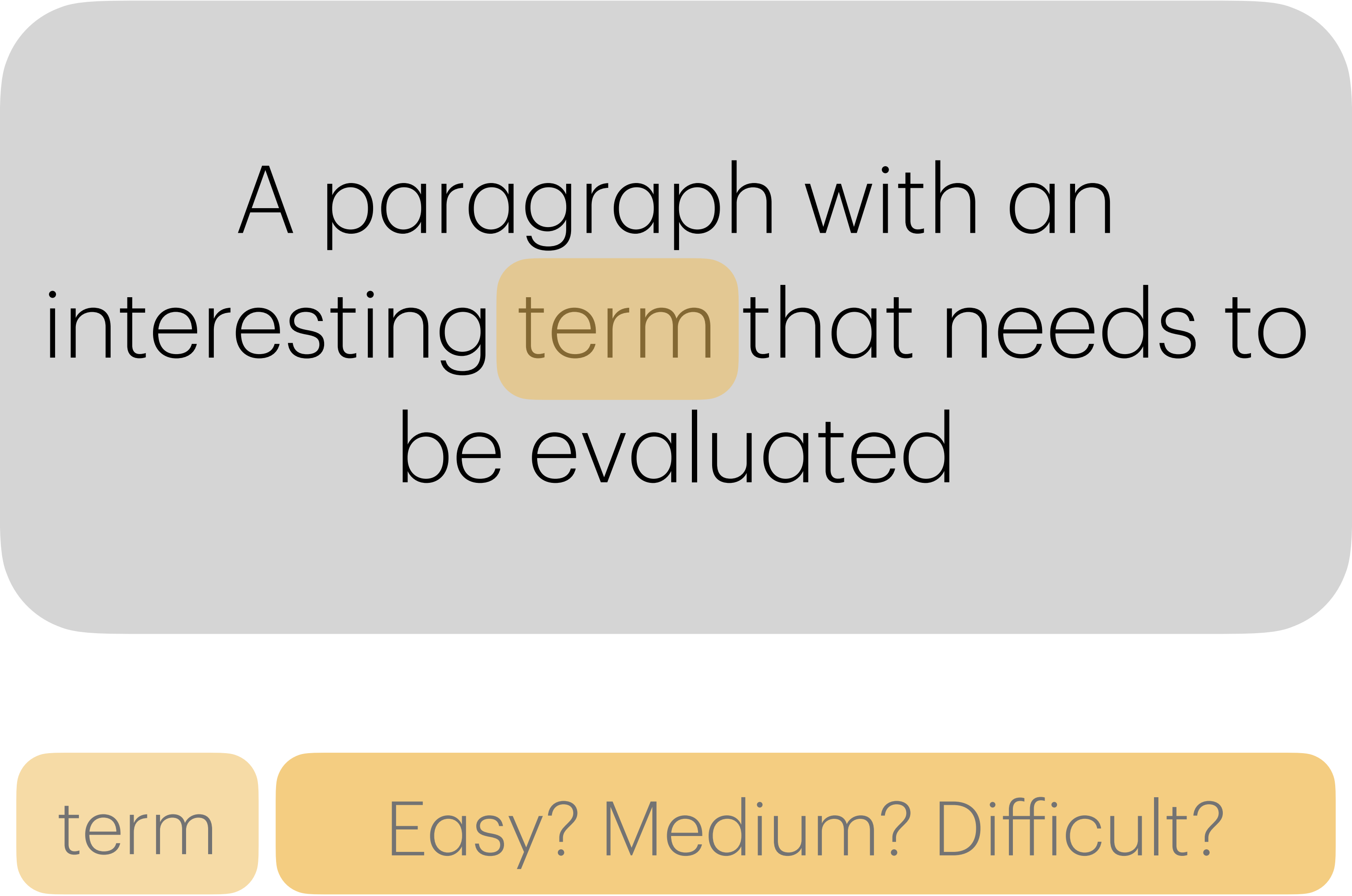
  - Task 2.1: predict the terms in a text and the difficulty of the concepts they designate (easy/medium/difficult).

  - Task 2.2: write a definition (and an explanation) for each difficult term.

  - Task 2.3: retrieve the provided definitions of the difficult terms in "correct" order.

Task 2.1

Find candidate terms and set difficulty



A paragraph with an interesting term that needs to be evaluated

term

Easy? Medium? Difficult?

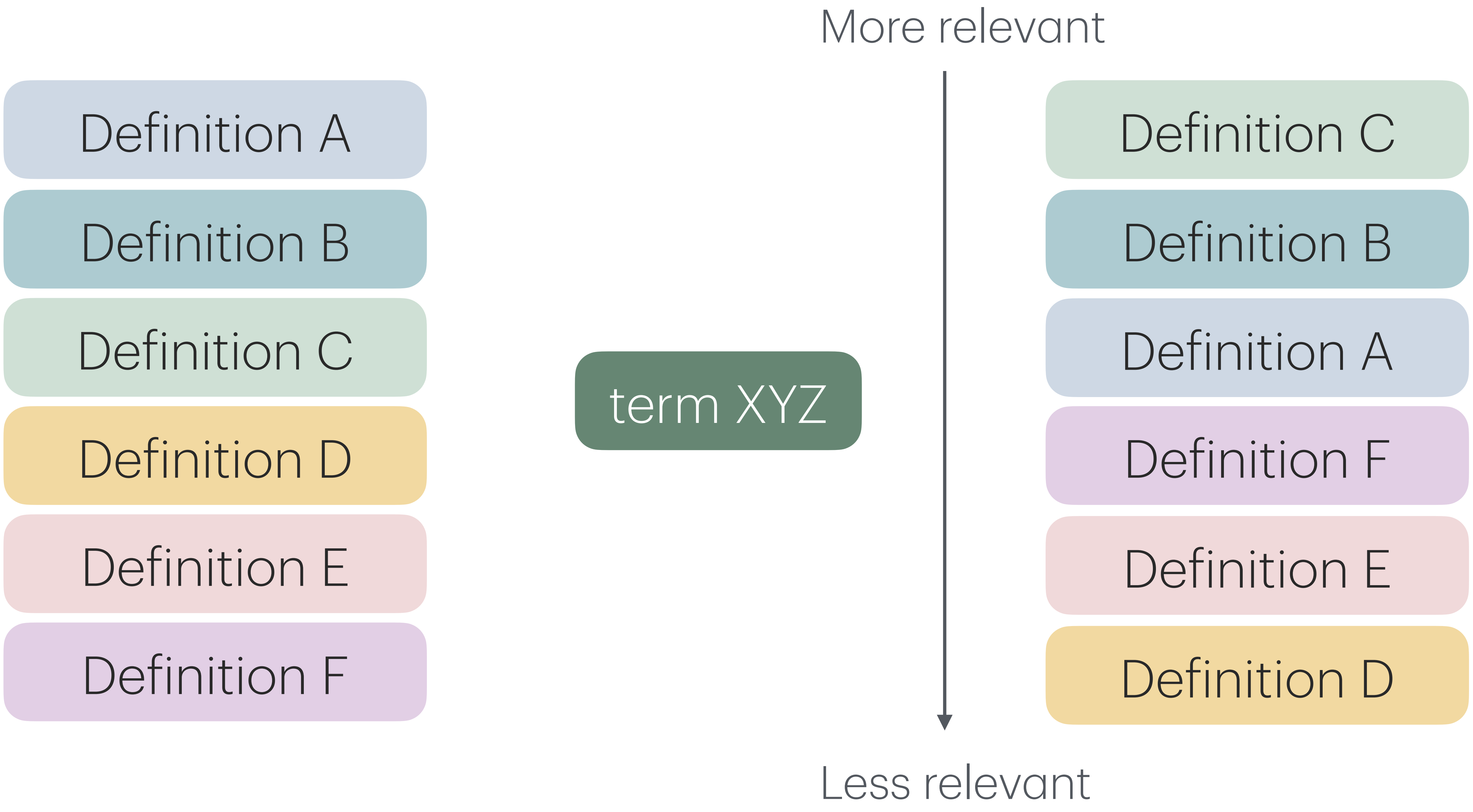
Task 2.2

Write definitions (and explanations) for difficult concepts

concept 1	term 1	Easy		
concept 2	term 2	Difficult	Definition	Explanation
concept 3	term 3	Medium		
concept 3	term 4	Difficult	Definition	Explanation

Task 2.3

Rank available definition



Definition A

Definition B

Definition C

Definition D

Definition E

Definition F

term XYZ

More relevant

Definition C

Definition B

Definition A

Definition F

Definition E

Definition D

Less relevant

Dataset creation

- The corpus of Task 2 is based on the sentences in high-ranked abstracts to the requests of Task 1. And collected in 2023.
- A total of 175 documents and 1,077 sentences were used to generate the training and test data. In particular, we had
  - 115 documents and 576 sentences for the training set and
  - 60 documents and 501 sentences for the test set

Dataset creation

- In particular, the dataset comprises the following files
  - The documents and their sentences.
  - Terms manually extracted and their relative difficulty.
  - Definitions and the explanations provided for the difficult terms.
  - Definitions automatically generated by a large language model.

Dataset creation

Training set

- For the training set, we engaged 21 experts to manually annotate each document, identifying the terms in each sentence, assessing their difficulty, and providing definitions and explanations for each difficult term.
- This effort resulted in the generation of 1,609 terms and 899 definitions and explanations.
- To further analyze the consistency among experts, we deliberately assigned the same documents to multiple experts in some instances.
- Additionally, for each concept accompanied by a definition, we created two "good" definitions and two "bad" definitions.

Dataset creation

Validation set

- We introduced an additional set of files produced by an external expert who reviewed the annotations of the 21 experts.
- This secondary set, referred to as the validation set, included the expert's additions of missing terms, definitions, or both.
- This review added 677 terms, 960 definitions.
- In addition 3,732 generated definitions (equally divided between good and bad) were added.

Dataset creation

Validation set

- For the test set, we asked the external expert to annotate the remaining 60 documents.

- A total of 1,440 terms were extracted and 424 definitions were written from the 501 sentences of the test set.

- An additional 3,816 definitions (equally distributed between good and bad definitions) were also added.

Dataset Annotation

Two steps

- Identification and manual extraction of candidate terms from abstracts
  - Term: “designation that represents a general concept by linguistic means” ISO 1087: 2019
- Construction of a collection of definitions of the concepts designated by candidate terms

Dataset Annotation

Creation of definitions

- Retrieval of definitions of the concepts from sources
- Transformation – where necessary – of definitions into intensional definitions
    - Intensional definition: “definition that conveys the intension of a concept by stating the immediate generic concept and the delimiting characteristic(s)” ISO 1087: 2019

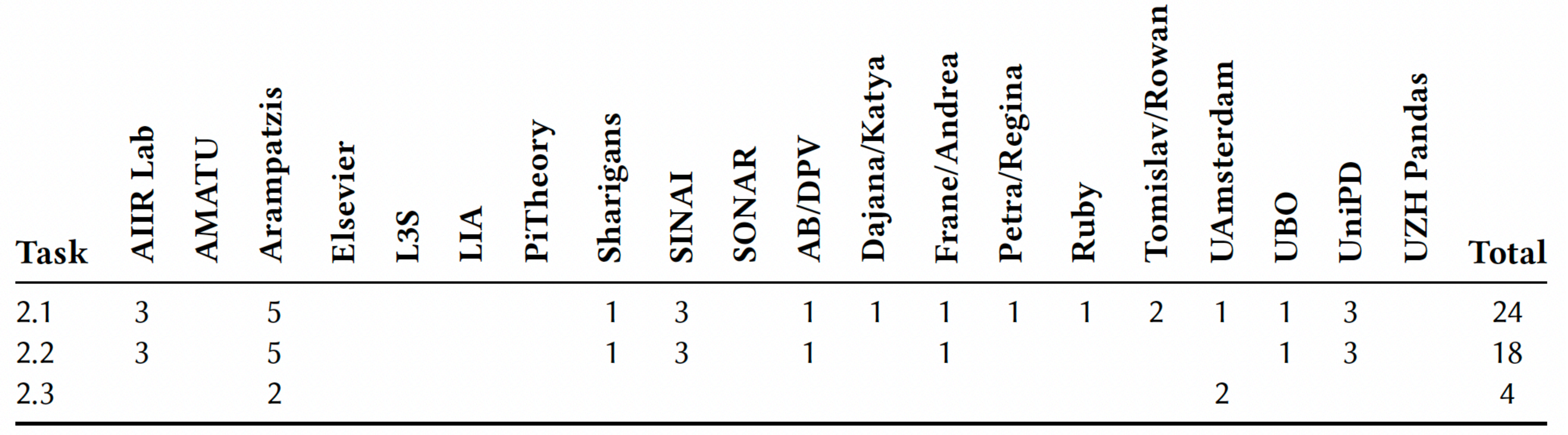
Evaluation`

Metrics

- Recall of all the terms, independently from the level of difficulty
- Precision of all the terms, independently from the level of difficulty
- • the F1 score of all the terms, independently from the level of difficulty
- Recall of the difficult terms
- Precision of the difficult terms
- F1 score of the difficult terms
- BLEU score computed for bigrams (ngrams from n = 1 to n = 4)

Results

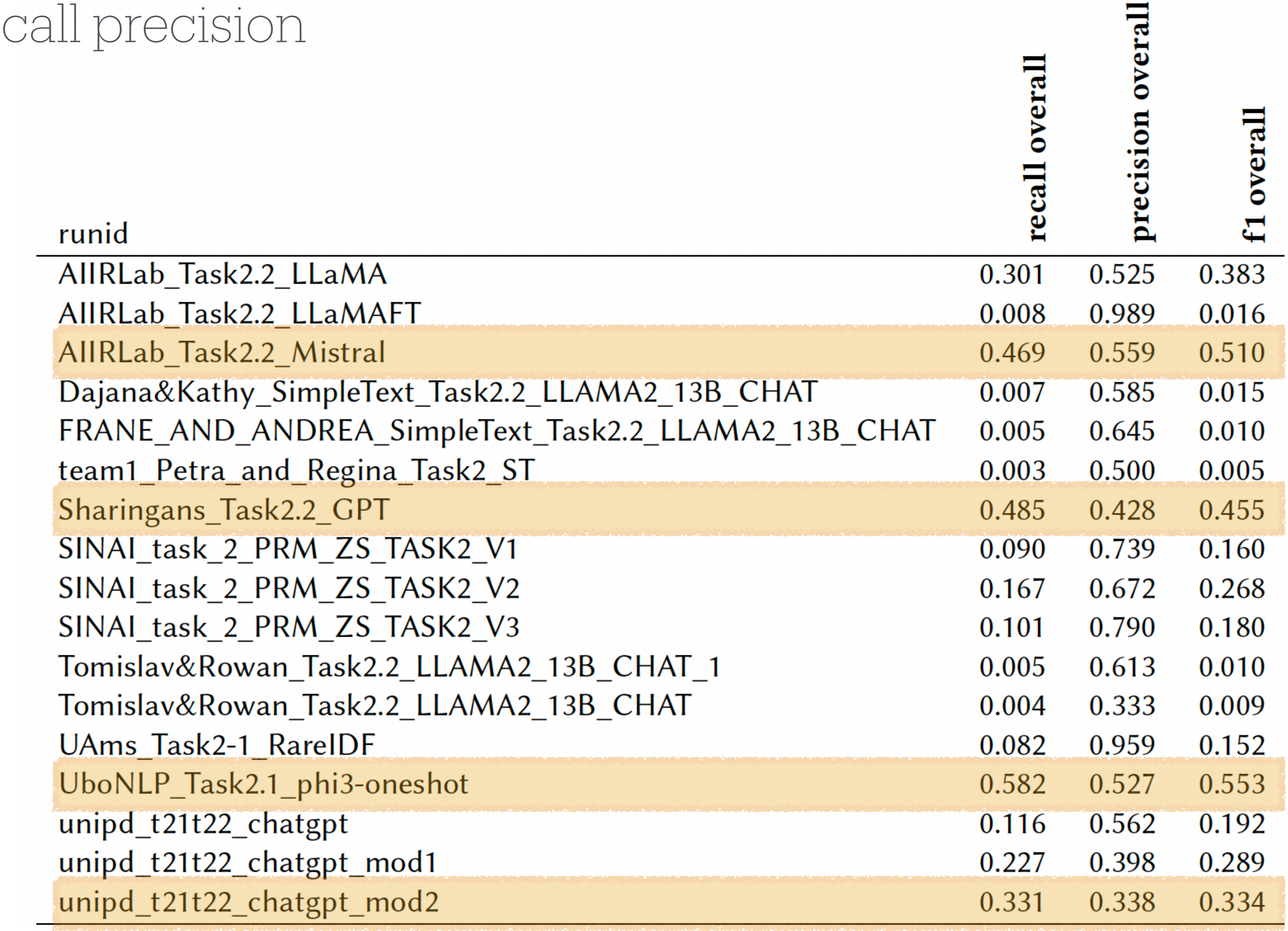
Participants



Task	AIIR Lab	AMATU	Arampatzis	Elsevier	L3S	LIA	PiTheory	Sharigans	SINAI	SONAR	AB/DPV	Dajana/Katya	Frane/Andrea	Petra/Regina	Ruby	Tomislav/Rowan	UAmsterdam	UBO	UniPD	UZH Pandas	Total
2.1	3		5					1	3		1	1	1	1	1	2	1	1	3		24
2.2	3		5					1	3		1		1					1	3		18
2.3			2														2				4

Results

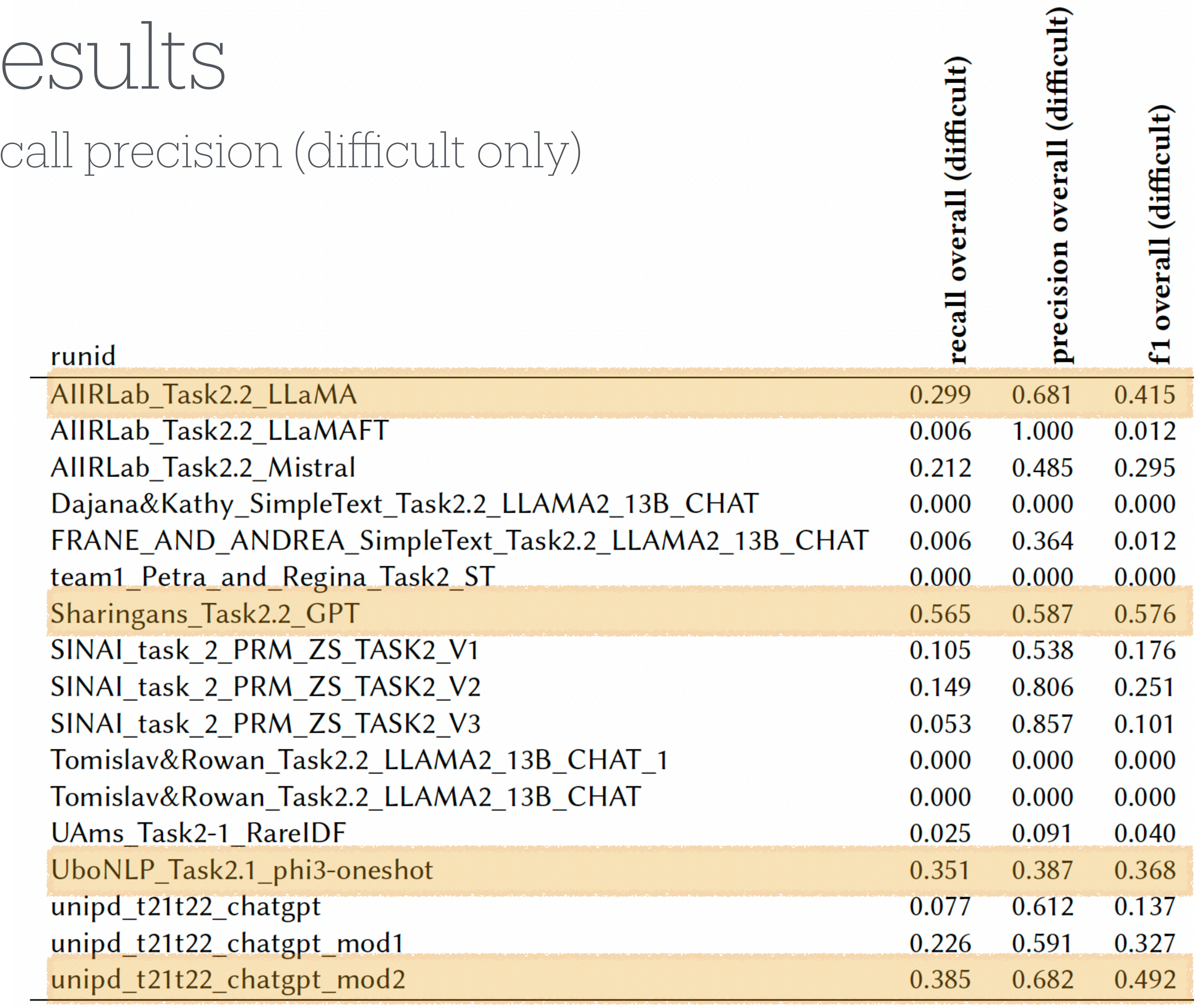
Recall precision



runid	recall overall	precision overall	f1 overall
AIIRLab_Task2.2_LLaMA	0.301	0.525	0.383
AIIRLab_Task2.2_LLaMAFT	0.008	0.989	0.016
AIIRLab_Task2.2_Mistral	0.469	0.559	0.510
Dajana&Kathy_SimpleText_Task2.2_LLAMA2_13B_CHAT	0.007	0.585	0.015
FRANE_AND_ANDREA_SimpleText_Task2.2_LLAMA2_13B_CHAT	0.005	0.645	0.010
team1_Petra_and_Regina_Task2_ST	0.003	0.500	0.005
Sharingans_Task2.2_GPT	0.485	0.428	0.455
SINAI_task_2_PRM_ZS_TASK2_V1	0.090	0.739	0.160
SINAI_task_2_PRM_ZS_TASK2_V2	0.167	0.672	0.268
SINAI_task_2_PRM_ZS_TASK2_V3	0.101	0.790	0.180
Tomislav&Rowan_Task2.2_LLAMA2_13B_CHAT_1	0.005	0.613	0.010
Tomislav&Rowan_Task2.2_LLAMA2_13B_CHAT	0.004	0.333	0.009
UAms_Task2-1_RareIDF	0.082	0.959	0.152
UboNLP_Task2.1_phi3-oneshot	0.582	0.527	0.553
unipd_t21t22_chatgpt	0.116	0.562	0.192
unipd_t21t22_chatgpt_mod1	0.227	0.398	0.289
unipd_t21t22_chatgpt_mod2	0.331	0.338	0.334

Results

Recall precision (difficult only)



runid	recall overall (difficult)	precision overall (difficult)	f1 overall (difficult)
AIIRLab_Task2.2_LLaMA	0.299	0.681	0.415
AIIRLab_Task2.2_LLaMAFT	0.006	1.000	0.012
AIIRLab_Task2.2_Mistral	0.212	0.485	0.295
Dajana&Kathy_SimpleText_Task2.2_LLAMA2_13B_CHAT	0.000	0.000	0.000
FRANE_AND_ANDREA_SimpleText_Task2.2_LLAMA2_13B_CHAT	0.006	0.364	0.012
team1_Petra_and_Regina_Task2_ST	0.000	0.000	0.000
Sharingans_Task2.2_GPT	0.565	0.587	0.576
SINAI_task_2_PRM_ZS_TASK2_V1	0.105	0.538	0.176
SINAI_task_2_PRM_ZS_TASK2_V2	0.149	0.806	0.251
SINAI_task_2_PRM_ZS_TASK2_V3	0.053	0.857	0.101
Tomislav&Rowan_Task2.2_LLAMA2_13B_CHAT_1	0.000	0.000	0.000
Tomislav&Rowan_Task2.2_LLAMA2_13B_CHAT	0.000	0.000	0.000
UAms_Task2-1_RareIDF	0.025	0.091	0.040
UboNLP_Task2.1_phi3-oneshot	0.351	0.387	0.368
unipd_t21t22_chatgpt	0.077	0.612	0.137
unipd_t21t22_chatgpt_mod1	0.226	0.591	0.327
unipd_t21t22_chatgpt_mod2	0.385	0.682	0.492

Results

BLEU (difficult terms only)

runid	BLEU (n1) average	BLEU (n2) average	BLEU (n3) average	BLEU (n4) average
AIIRLab_Task2.2_LLaMA	0.286	0.150	0.047	0.018
AIIRLab_Task2.2_LLaMAFT	0.240	0.117	0.000	0.000
AIIRLab_Task2.2_Mistral	0.259	0.133	0.041	0.014
Dajana&Kathy_SimpleText_Task2.2_LLAMA2_13B_CHAT	0.000	0.000	0.000	0.000
FRANE_AND_ANDREA_SimpleText_Task2.2_LLAMA2_13B_CHAT	0.000	0.000	0.000	0.000
team1_Petra_and_Regina_Task2_ST	0.000	0.000	0.000	0.000
Sharingans_Task2.2_GPT	0.227	0.106	0.031	0.016
SINAI_task_2_PRM_ZS_TASK2_V1	0.252	0.157	0.082	0.060
SINAI_task_2_PRM_ZS_TASK2_V2	0.276	0.159	0.067	0.049
SINAI_task_2_PRM_ZS_TASK2_V3	0.216	0.112	0.039	0.025
Tomislav&Rowan_Task2.2_LLAMA2_13B_CHAT_1	0.000	0.000	0.000	0.000
Tomislav&Rowan_Task2.2_LLAMA2_13B_CHAT	0.000	0.000	0.000	0.000
UAms_Task2-1_RareIDF	0.000	0.000	0.000	0.000
UboNLP_Task2.1_phi3-oneshot	0.001	0.000	0.000	0.000
unipd_t21t22_chatgpt	0.309	0.185	0.089	0.049
unipd_t21t22_chatgpt_mod1	0.311	0.181	0.082	0.045
unipd_t21t22_chatgpt_mod2	0.294	0.184	0.091	0.052

Conclusion

- Big effort to create
  - one of the largest dataset for the evaluation of Automatic Term Extraction (ATE) tools
  - First dataset for ATE with with data about the difficulty of a term
  - First dataset of terminological definitions
- Very interesting results
  - Many cases of human-in-the-loop with LLM

Conclusion

- Lots of things to do
    - Additional round of cleaning and re-evaluation of the dataset
    - Additional thoughts about the evaluation metrics
    - Better description for Task 2.3 (potentially a big impact)

SimpleText Task 2

Identify and Explain Difficult Concepts

Giorgio Maria Di Nunzio, Federica Vezzani, Vanessa Bonato, Hosein Azarbonyad,
Jaap Kamps and Liana Ermakova

9 September 2024, CLEF 2024, Grenoble, France