Advacheck at GenAI Detection Task 1
AI Detection Powered by Domain-Aware Multi-Tasking

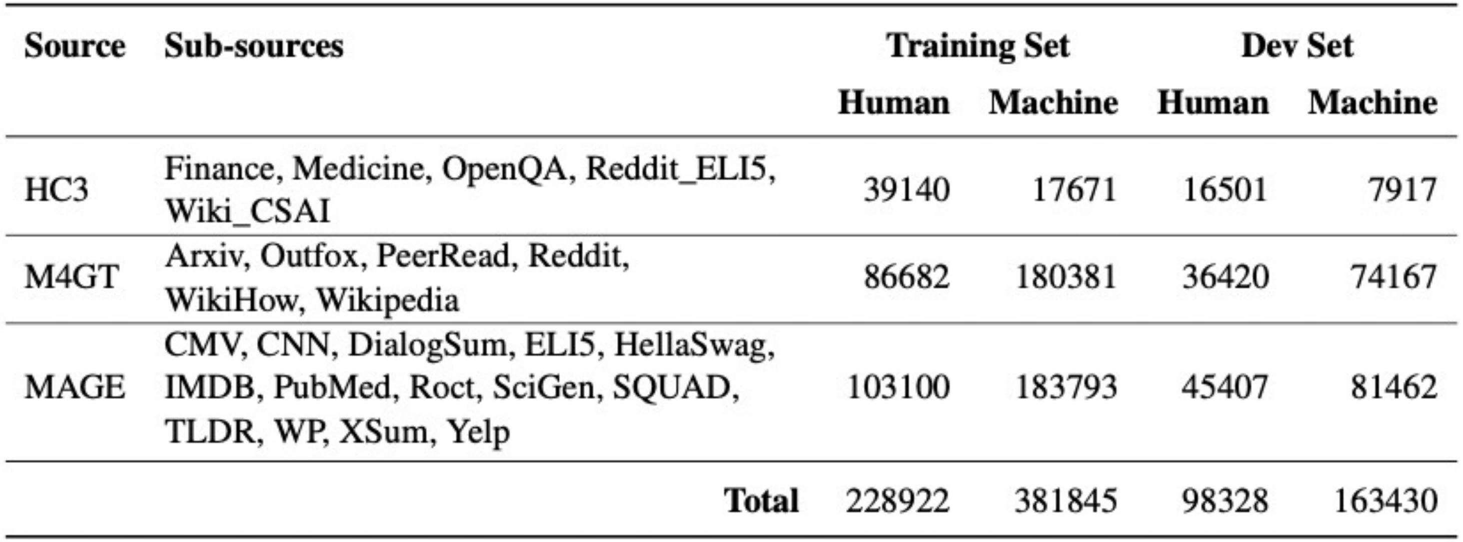
German Gritsai, Anastasia Voznyuk, Ildar Khabutdinov, Andrey Grabovoy

Advacheck Company
University Grenoble Alpes
{gritsai, voznyuk}@advacheck.com

19.01.2025

Motivation

Task participated: Subtask A - Binary Machine-Generated Text Detection
Data language: English
Metric: Macro F_1-score



Source	Sub-sources	Training Set		Dev Set	
		Human	Machine	Human	Machine
HC3	Finance, Medicine, OpenQA, Reddit_ELI5, Wiki_CSAI	39140	17671	16501	7917
M4GT	Arxiv, Outfox, PeerRead, Reddit, WikiHow, Wikipedia	86682	180381	36420	74167
MAGE	CMV, CNN, DialogSum, ELI5, HellaSwag, IMDB, PubMed, Roct, SciGen, SQUAD, TLDR, WP, XSum, Yelp	103100	183793	45407	81462
	Total	228922	381845	98328	163430

Table. Statistics on training and development data from monolingual subtask of the GenAI Detection Task 1.

3 sources
25 sub-sources
> 600k train data
40 models

Model Architecture

Binary CCH – 2 classes:

- Initial monolingual statement

Multiclass CCH – 5 classes:

- Sub-source within HC3

Multiclass CCH – 6 classes:

- Sub-source within M4GT

Two-stage training:

- 🥶 fine-tuning classifiers with frozen shared encoder weights
- 🔥 fine-tuning the complete model with all weights unfrozen

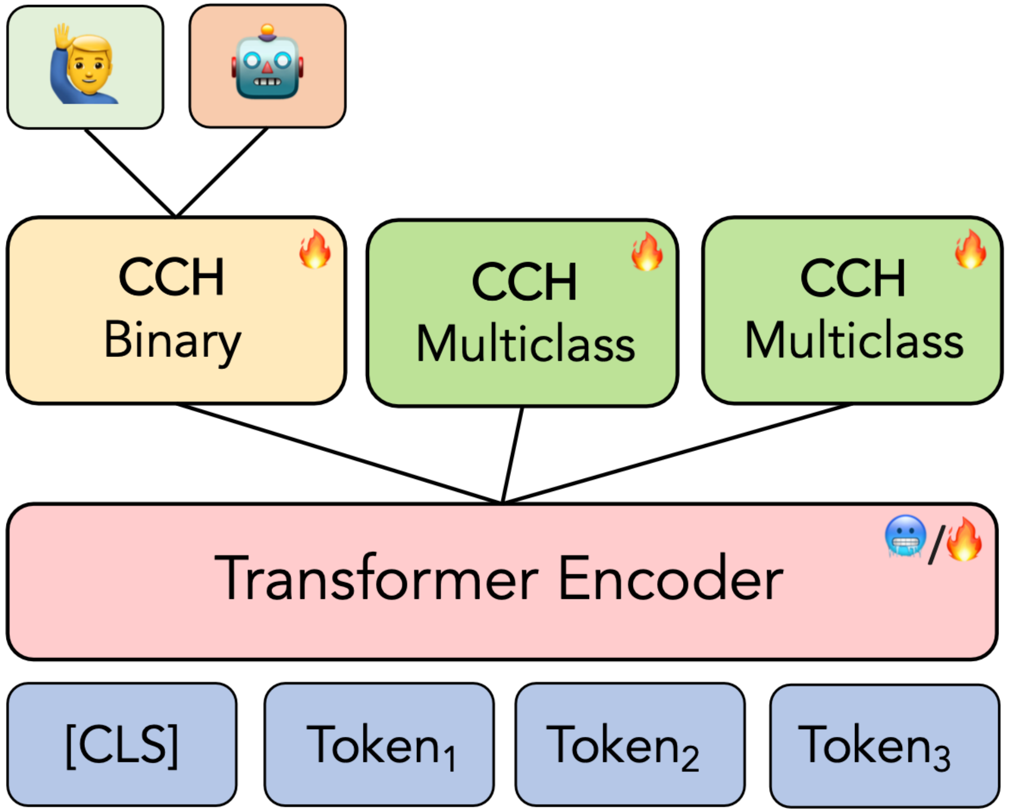
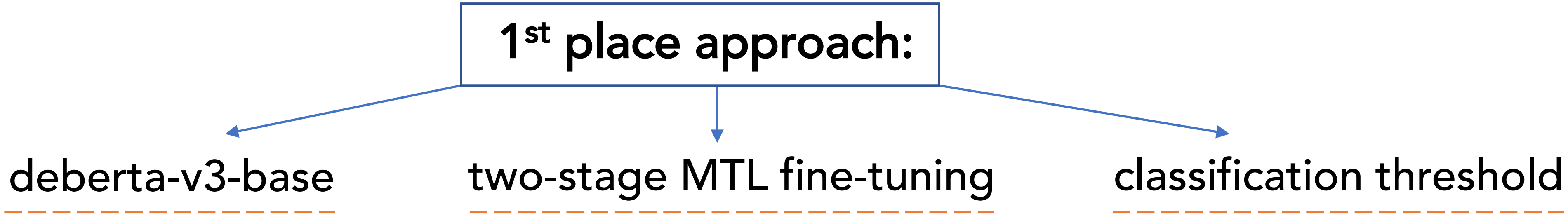


Figure. Proposed multi-task architecture with hard parameter sharing. CCH – Custom Classification Head.

Configuration Comparison

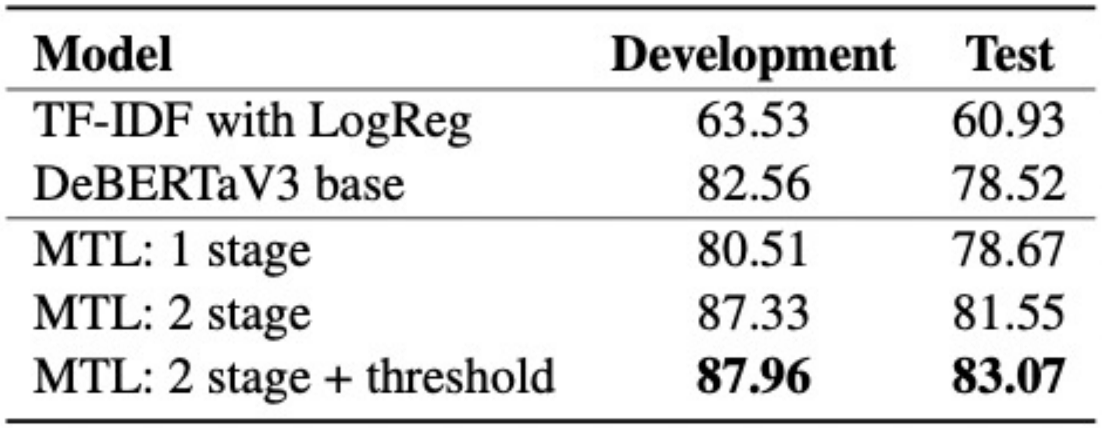


1st place approach:

deberta-v3-base

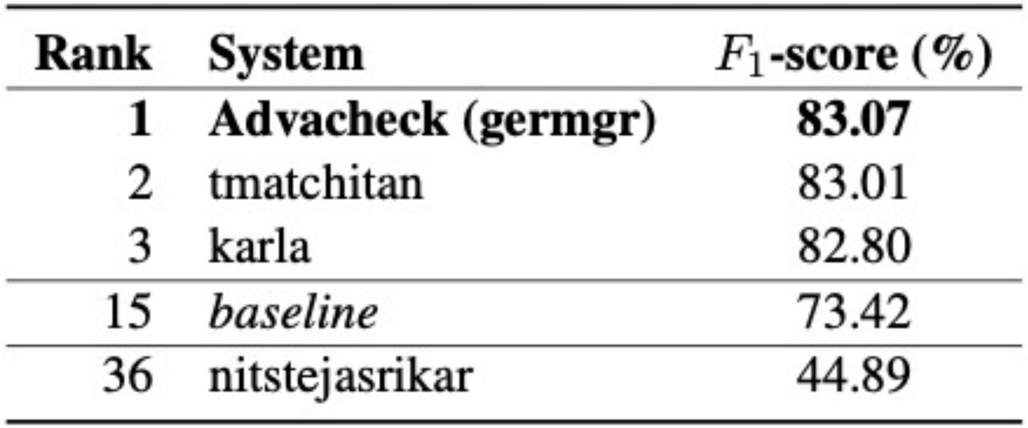
two-stage MTL fine-tuning

classification threshold



Model	Development	Test
TF-IDF with LogReg	63.53	60.93
DeBERTaV3 base	82.56	78.52
MTL: 1 stage	80.51	78.67
MTL: 2 stage	87.33	81.55
MTL: 2 stage + threshold	**87.96**	**83.07**

Table: Results of model comparison on the test and development set.

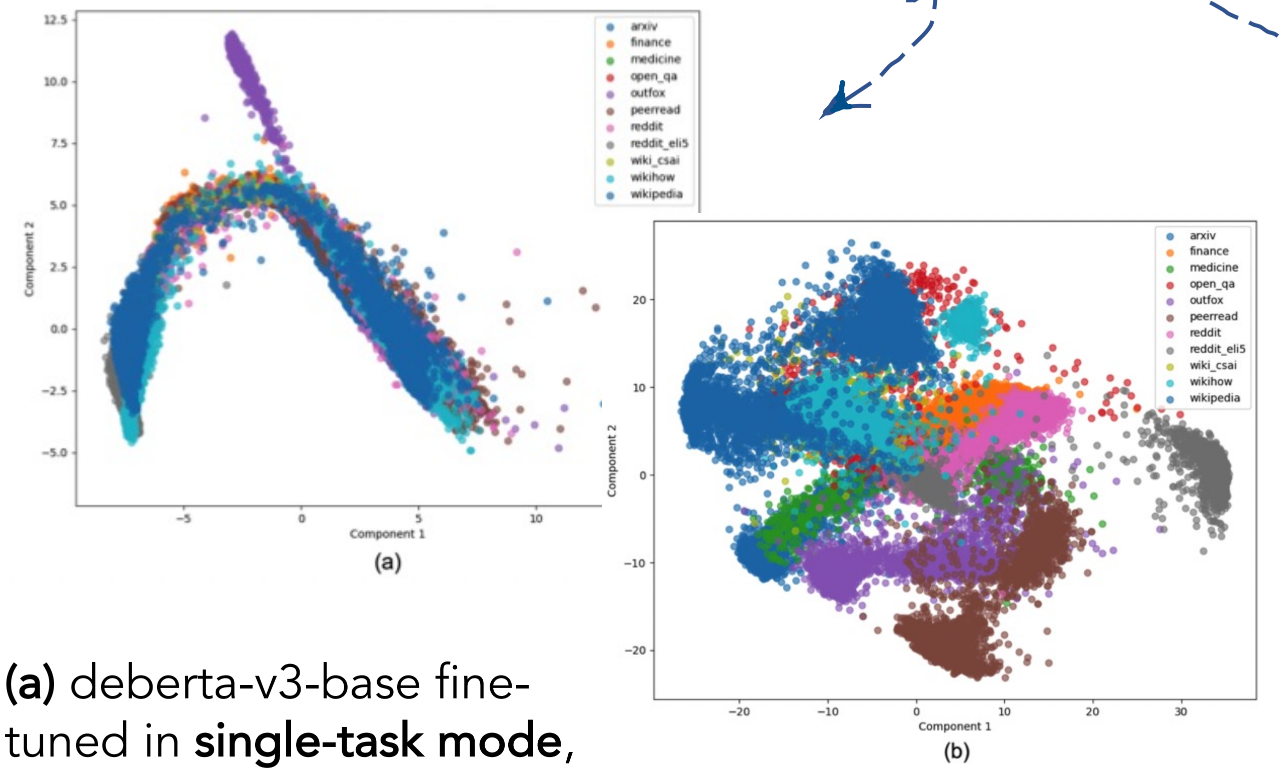


Rank	System	F_1-score (%)
1	**Advacheck (germgr)**	**83.07**
2	tmatchitan	83.01
3	karla	82.80
15	*baseline*	73.42
36	nitstejasrikar	44.89

Table: Final results on the official ranking. Bold denotes our team's placement.

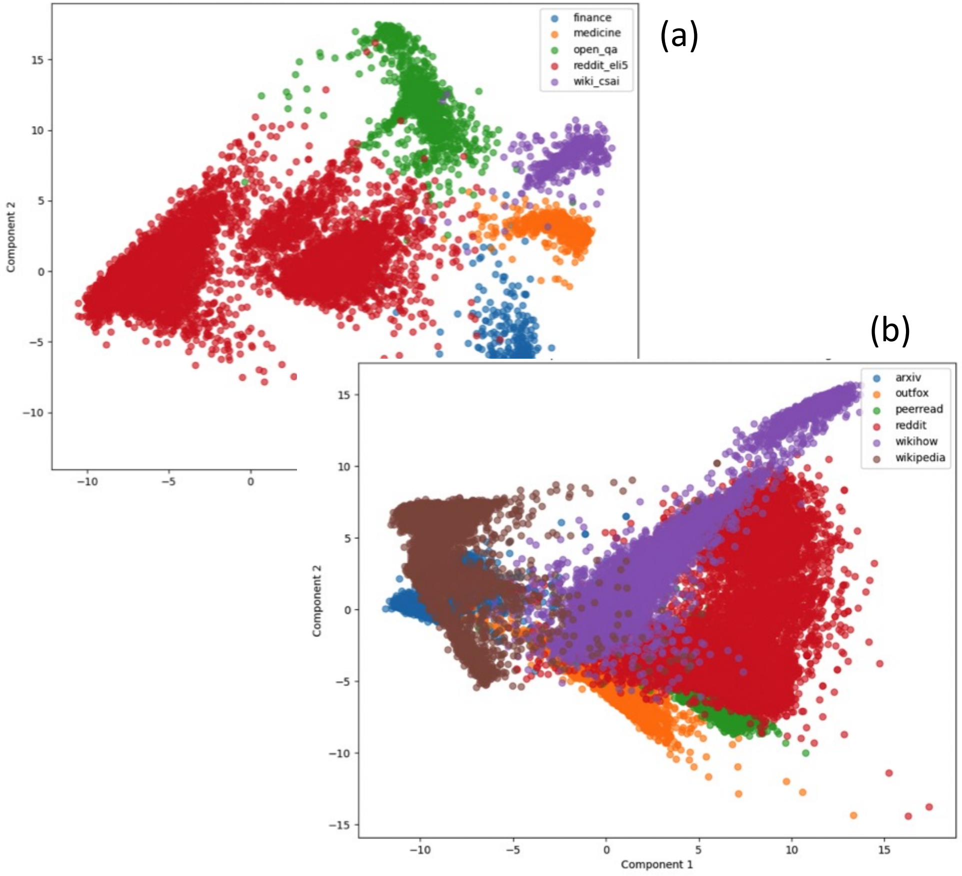
Embeddings after Multi-Task Learning

PCA decomposition for texts from the development subsample



(a) deberta-v3-base fine-tuned in **single-task mode**,

(b) the same model but fine-tuned in **MTL mode**.

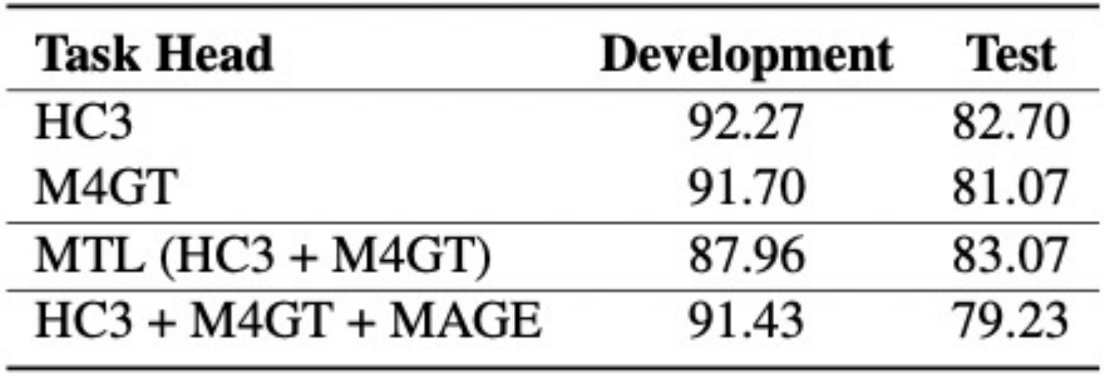


Logit decomposition of two multiclass CCH **(a) HC3** and **(b) M4GT**.

Ablations

Variation of multiple components of the system:

- **Heads with their quantity**
- **Threshold**



Task Head	Development	Test
HC3	92.27	82.70
M4GT	91.70	81.07
MTL (HC3 + M4GT)	87.96	83.07
HC3 + M4GT + MAGE	91.43	79.23

Table. Comparison of different configurations of heads and tasks trained simultaneously in MTL architecture.

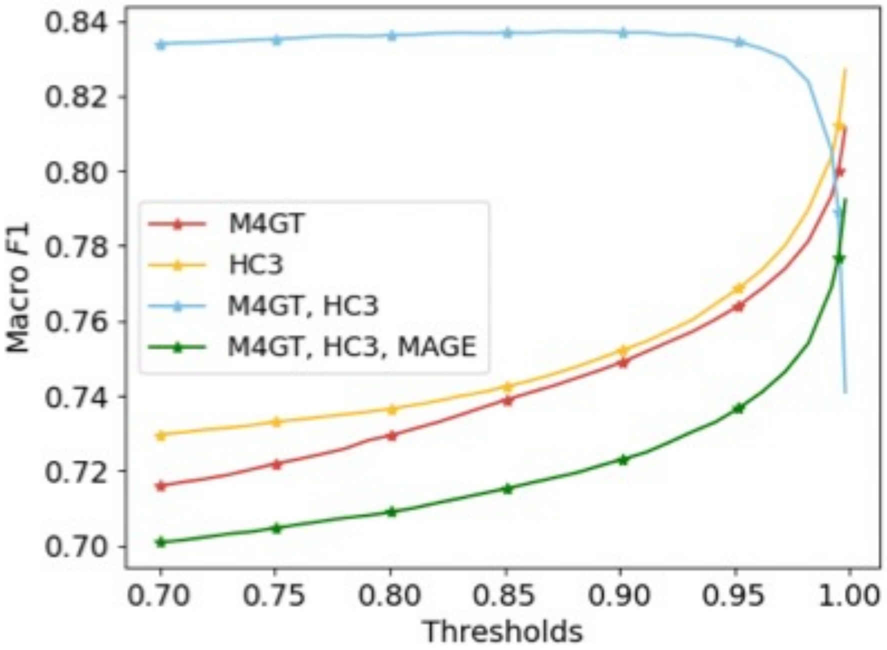


Figure. Macro F_1-score on the test set of different configuration of the systems depending on the threshold.

Error Analysis

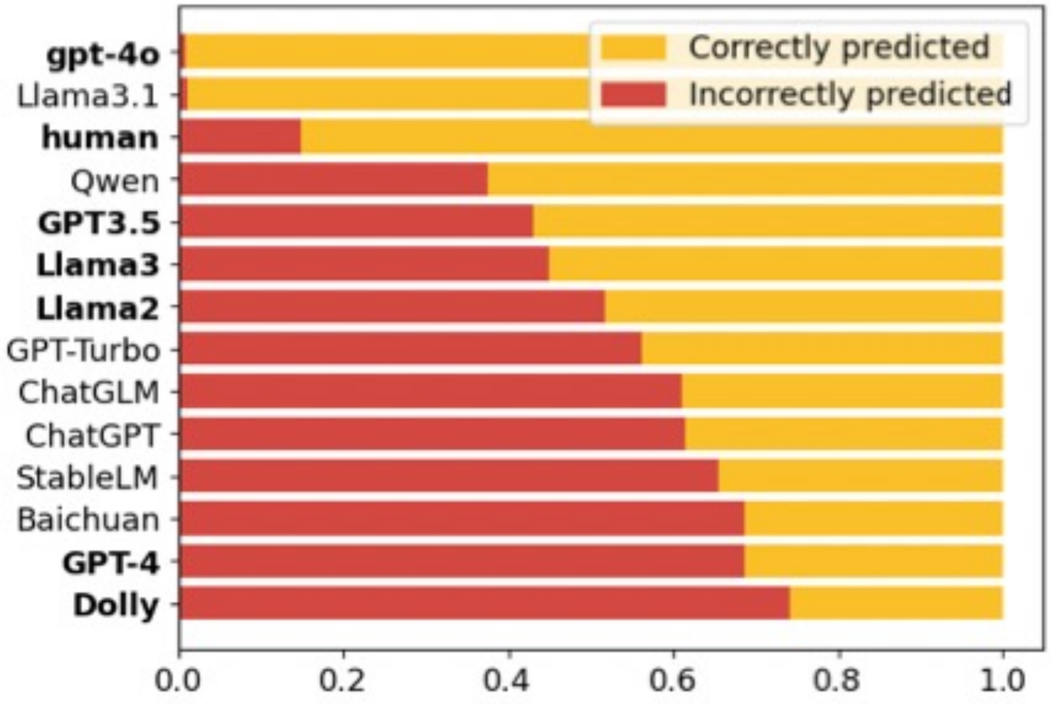


Figure. Proportion of predictions for different generators from test set. Labels in bold are generators texts from which are present in the train set.

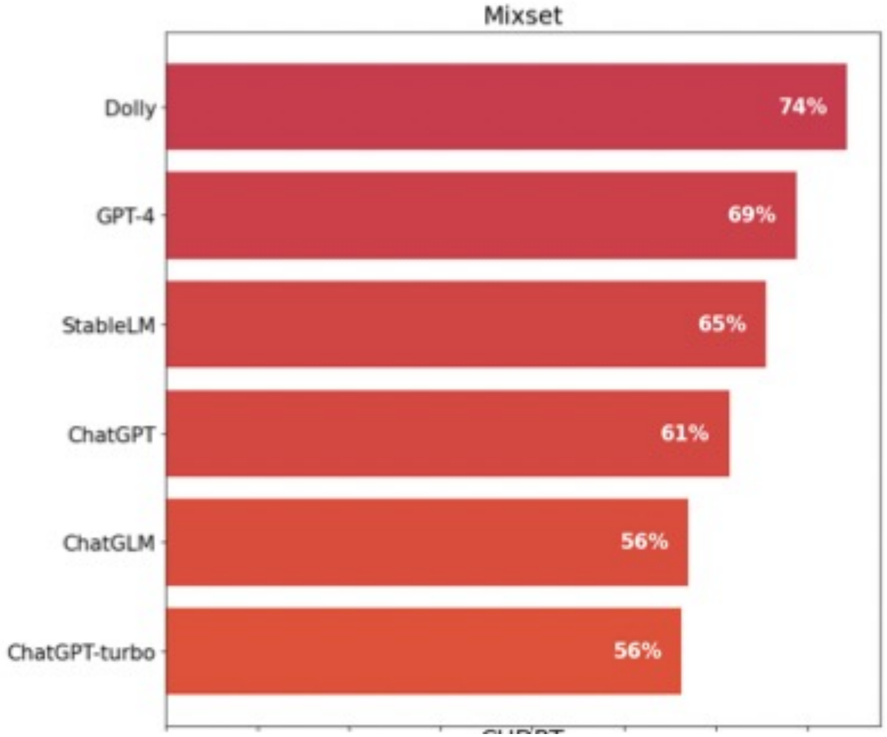


Figure. Dataset from test set with the highest percentage of incorrect predictions. There were some **additional manipulations with texts** after generation.

To conclude

Our winning approach:

deberta-v3-base as shared encoder

two-stage MTL fine-tuning forms cluster-wise structure

classification threshold

Future work:

- MTL as regularization
- Improve robust to the change of generators

Contact us:

"Are AI Detectors Good Enough?"

Advacheck at GenAI Detection Task 1
AI Detection Powered by Domain-Aware Multi-Tasking

German Gritsai, Anastasia Voznyuk, Ildar Khabutdinov, Andrey Grabovoy

Advacheck Company
University Grenoble Alpes
{gritsai, voznyuk}@advacheck.com

19.01.2025