DeSegMa-IT at EVALITA 2026: Overview of the "Detection and Segmentation of Machine Generated Text in Italian" Task

Giovanni Puccetti[1,*,†], Andrea Pedrotti[1,†] and Andrea Esuli[1]

[1]Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo" (CNR-ISTI), Via G. Moruzzi 1, Pisa, 56124, Italy

Abstract

DeSegMa-IT's shared tasks aim to test the robustness of machine-generated text (MGT) detectors by evaluating their performance under settings where the IID assumption does not hold. While state-of-the-art MGT detectors report high accuracy, such results often rely on unrealistic experimental settings: for example, relying on prior knowledge of the text generator, or failing to consider domain shifts and efficient fine-tuning - or post-tuning - strategies. In DeSegMa-IT, participants are challenged with two sub-tasks: *(i)* document-level detection of MGTs and the *(ii)* human-machine text segmentation. This paper describes the released dataset, discusses the systems submitted by participants, and provides an initial analysis of the obtained results.

Keywords

Machine-Generated Text Detection, Text Segmentation, Text Classification

1. Introduction and Motivation

Recent advancements in Generative AI and Large Language Models (LLMs) have led to the development of systems, such as GPT-4 [1], Claude[1], Llama-3 [2] and DeepSeek-V3 [3] among others, that can generate text that is often indistinguishable from human-written content [4].

This capability, along with the many beneficial applications of LLMs, also enables malicious actors to create Machine Generated Text (MGT) for deceptive purposes. For example, it can be used to manipulate online traffic and spread misinformation through content farms [5] or to influence human revisions of sensitive documents in critical domains, such as scientific peer review.[2]

Beyond malicious use, the widespread adoption of LLMs into everyday tools raises concerns related to authorship attribution, intellectual property, and the transparency of human-AI collaboration. In journalistic [5], educational [6], and governmental settings [7], the ability to distinguish between human-written and machine-generated content has become of uttermost importance to preserve the essential features of trust for such sensible domains, while also allowing for responsible AI deployment.

As a consequence, the task of machine-generated text (MGT) detection has received increasing attention throughout the years. With the rapid proliferation of AI-based assistants, the need for automatic tools to detect their outputs has become more urgent, leading to the proposal of numerous detection methods [8, 9, 10]. However most existing work focuses on English, highlighting the importance to develop MGT detection systems for other languages as well. Furthermore, the widespread availability of open-weight LLMs poses existing detection methods with the challenge of an ever-growing array of fine-tuned models, making the detection task increasingly non-IID (independent and identically distributed) and stressing their generalization capabilities to account for subtle shifts in writing style [11].

To address this gap, the community proposed shared tasks focused on the detection of MGT texts mainly focused on English or on multilingual settings [12, 13] along with the release of benchmark

EVALITA 2026: 9th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, Feb 26 – 27, Bari, IT

*Corresponding author.

†These authors contributed equally.

giovanni.puccetti@isti.cnr.it (G. Puccetti); andrea.pedrotti@isti.cnr.it (A. Pedrotti); andrea.esuli@isti.cnr.it (A. Esuli)

[1]anthropic.com/claude-model-card

[2]aclweb.org/genai-peerreview-guidelines

Sub-Task A Example

Human Text: Viktor Orban, da quando il leghista è diventato ministro dell'Interno, si mandano segnali di apprezzamento reciproco. Un'alleanza che potrebbe portare l'Italia al fianco dei Paesi di Visegrad e che già - in occasione della discussione sulla riforma di Dublino - ha messo in difficoltà gli altri partner dell'Ue. Il primo ad allacciare rapporti era stato Salvini. Da Frosinone il ministro aveva parlato dell'Ungheria come di un paese con cui l'Italia potrà cambiare l'Europa. Entrambi, in fondo, sono dichiaratamente euroscettici. E sia Orban che il leghista ... **Label:** 0

Machine Text: Viktor Orban, dopo anni di duri scontri diplomatici, sono pronti a unire le loro forze per riscrivere l'agenda di Bruxelles. Il leader di Fratelli d'Italia Giorgia Meloni e il presidente del governo ungherese si sono visti a Vienna, in un vertice di "centrodestra, identità italiana e sovranità italiana". Dopo un incontro di ben tre ore i due hanno spiegato come si possano conciliare politicamente le due visioni d'Europa ... **Label:** 1

Figure 1: Sub-task A Example.

Sub-task B Example

Text: Il presidente del Tribunale internazionale del diritto del mare (Itlos), Vladimir Golitsyn, ha fissato *al 10 agosto la data in cui il tribunale arbitrale di Amburgo esaminerà le informazioni che l'Italia intende raccogliere in India per scagionare i Marò . Nei giorni scorsi su vari quotidiani erano uscite indiscrezioni circa la data di un eventuale incontro tra i due fucilieri di Marina e i loro avvocati e i funzionari del ministero dell'Interno indiano, che dovrebbero rilasciare a loro una sorta di "licenza" temporanea così che i due marinai possano recarsi in India.*

Target Character Index: 103

Figure 2: Sub-task B Example. Non-italic text represents the human segment, while the *italic* text denotes the continuation generated by a LLM.

datasets [14]. To extend these efforts to the Italian language, the EVALITA 2026 [15] shared task DeSegMa-IT aims to foster research on Italian MGTs by providing the research community with a benchmark dataset and a standardized evaluation framework.

2. Tasks Definitions

DeSegMa-IT[3] shared task is organized into two sub-tasks: *(i)* MGT detection and *(ii)* human–machine text segmentation. The first sub-task evaluates detection accuracy at the document level and aims to test the robustness of machine-generated text detectors under realistic, non-IID conditions. While state-of-the-art MGT detectors report high accuracy, such results often rely on unrealistic experimental settings—for example, depending on prior knowledge of the text generator (e.g., [16]) or ignoring domain shifts and fine-tuning strategies. To address this, we simulate real-world non-IID conditions by generating the training and testing datasets from different LLMs.

In contrast, the segmentation sub-task focuses on identifying machine-generated text within a human-written document. For this task, both dataset splits share the same LLM generators due to its inherently higher difficulty.

2.1. Sub-task A: MGT Detection in the Wild

In **sub-task A**, we simulate the challenge posed by the ever-shifting domain of MGT detection by sampling train and test documents from two disjunct sets of generating LLMs. The task is structured as a binary-classification problem and defined as follows: *"Given a piece of text t, assign it the label 0, if the text is written by a human, and 1 otherwise."* We provide an example in Figure 1.

[3] https://desegma.github.io/

Source	Prompt
Il Giornale	Sei un giornalista italiano che scrive sul giornale conservatore di destra "Il Giornale". Scrivi un articolo di giornale a partire da questo titolo: <NEWS TITLE>. Evita qualsiasi tipo di formattazione. Non generare il titolo, inizia direttamente dal corpo dell'articolo.
La Repubblica	Sei un giornalista italiano che scrive sul giornale progressista di sinistra "La Repubblica". Scrivi un articolo di giornale a partire da questo titolo: <NEWS TITLE>. Evita qualsiasi tipo di formattazione. Non generare il titolo, inizia direttamente dal corpo dell'articolo.

Table 1
Prompts used for the generations.

2.2. Sub-task B: Human - Machine Text Segmentation

In the **second sub-task**, participants are required to detect the boundary between the human-written text and the machine-generated continuation by identifying the index of the character that marks the beginning of the MGT content. Each data sample consists of a variable-length human-written prompt, always followed by a variable-length continuation produced by the model. Unlike traditional MGT detection tasks that require document-level binary classification, this sub-task focuses on segmentation: participants must pinpoint the beginning of the text generated by the LLM.

The task is defined as follows: *"Given a piece of text t, return the index of the first character that is generated by an LLM."* To ensure a statistically robust evaluation, the length of the human-written sub-string is uniformly sampled from a range of 64 to 512 characters. This setup simulates real-world scenarios in which MGT may be inserted into otherwise human-written content. We provide an example in Figure 2.

3. Dataset

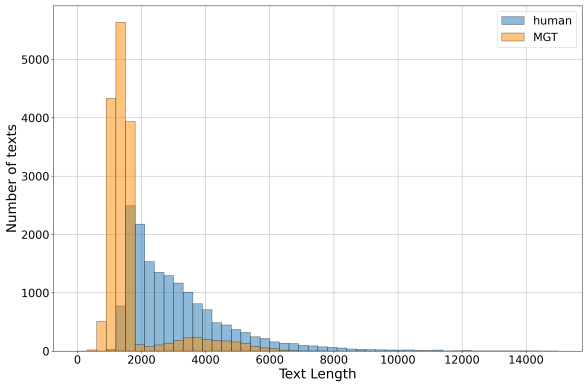
For each of the two sub-tasks, we provide participants with training and evaluation specific data. In this section, we describe the two datasets in details.

3.1. Sub-task A: MGT Detection

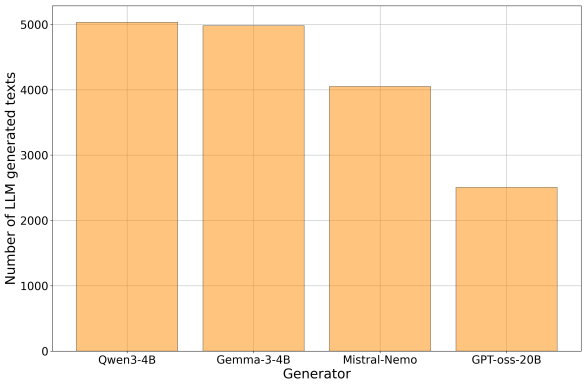
The dataset for sub-task A consists of both human-written and machine-generated texts. Human texts are sampled from the Change-IT dataset [17]. The original dataset consists of news articles collected from two Italian outlets: the *La Repubblica*[4] newspaper and the *Il Giornale*[5] newspaper. From the dataset, we retain the headline field, storing the title of the news articles and the human-written article itself. We use the headline to prompt a pool of LLMs to generate an synthetic version of the article. Furthermore, we provide the LLM with guidelines regarding the political agenda of the original news outlet. We also prompt the model to avoid any formatting style (e.g., retain from using bullet points, markdown sections, etc.) to adhere more closely to the style of the human-written articles. All models are prompted with their respective default generation parameters. We report the prompts used in Table 1. We sample a random balanced selection of human-written and machine-generated texts to construct the task' datasets.

Training dataset We collect 33,138 news articles, equally split between human and machine-generated texts, while preserving a 1:1 ratio for each news outlet. The models used to create the machine generated part of the training dataset are reported in Appendix A.1.

[4]https://www.repubblica.it/
[5]https://www.ilgiornale.it/

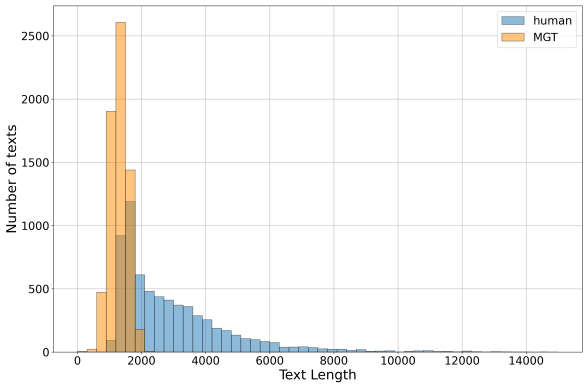


(a) Text length distribution by class (training set)

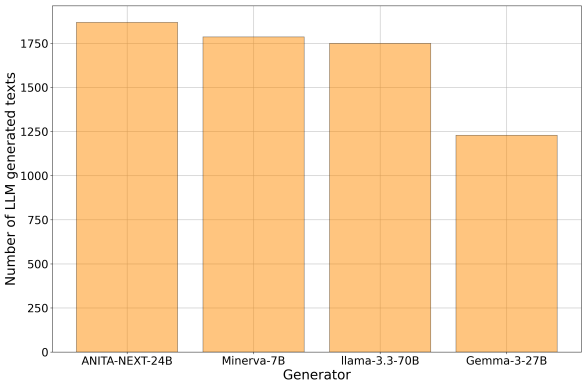


(b) Number of texts per generator (training set)

Figure 3: Text statistics for sub-task A training data.



(a) Text length distribution by class (test set)



(b) Number of texts per generator (test set)

Figure 4: Text statistics for sub-task A test data.

We sample texts from the generators to maintain a balanced distributions among LLMs. Figure 3a reports the distribution of text lengths and Figure 3b shows the generating LLMs distribution. We include 15,135 articles, balanced according to the target classes and evenly split among generators.

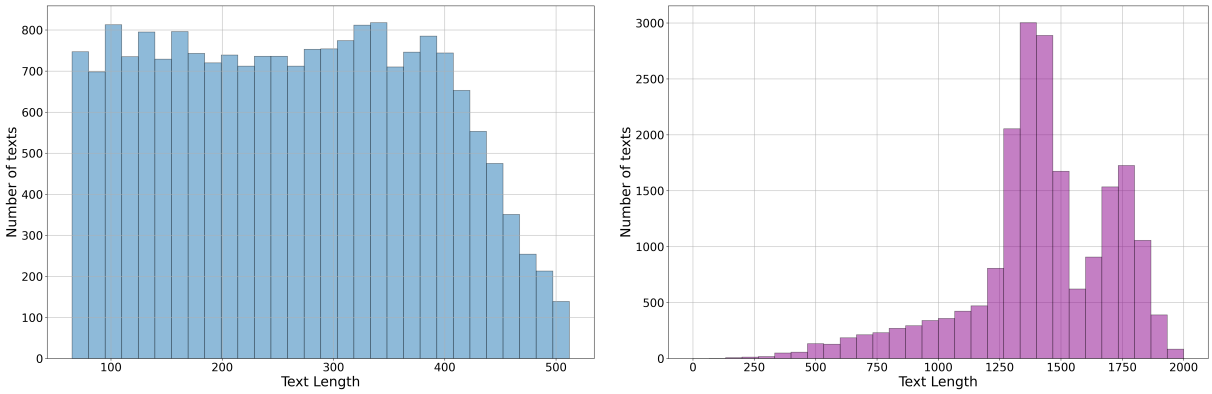
Test dataset We include 13,268 articles evenly balanced across news outlets. Figure 4a reports text length statistics and Figure 4b shows the distribution of texts with respect to the generating LLMs. Selected generators are reported in Appendix A.1.

3.2. Filtering by Perplexity

To make the classification more challenging, we compute with a held-out model, `gemma-3-12b-it`, the mean perplexity score for the human-written texts. We use this mean value to filter out MGTs that drift too far away from the selected filtering score. This process allows us to select documents that are similar in style according to the evaluator LLM, shrinking the classification boundary between the two classes.

3.3. Sub-task B: Human-MGT Segmentation

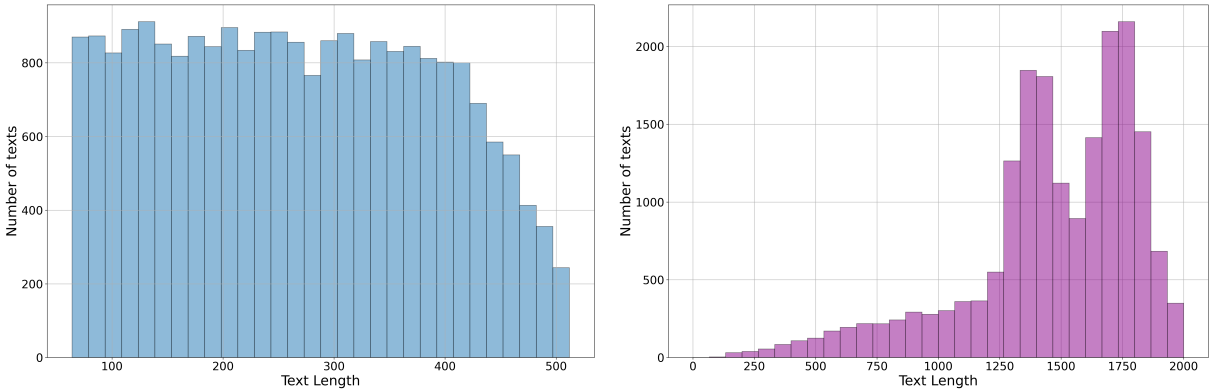
Data for the sub-task B consist of news articles switching from human written content to machine-generated continuation. Each text is paired with an integer denoting the length of the human written part. To generate plausible continuations, we select a human-written article and discard up to the first three sentences of the text. The remaining segment is used as an input prompt and fed to one of nine LLMs. The rationale for discarding the initial portion of the article is to present participants with a more



(a) Human text length distribution

(b) Total text length distribution

Figure 5: Text length statistics for sub-task B training data.



(a) Human text length distribution (test set)

(b) Total text length distribution (test set)

Figure 6: Text length statistics for sub-task B test data.

challenging scenario, designed to mimic the detection of machine-generated text segments appearing at arbitrary positions within an article, rather than only at its beginning.

Training dataset This split consists of 19,945 news articles. Generating LLMs are reported in Appendix A.2. We set the minimum length of the human-written to 64 characters, and the maximum length to 512 characters. The length distribution of human-written segments is shows in Figure 5a and the length distribution of full human-written and machine-continued texts is reported in Figure 6b.

Test dataset For the test set, we keep the same LLMs and select 23,211 new samples, while maintaining an even distribution among models. Figure 6a shows the length distribution of the human segments at the beginning of each sample and Figure 6b the length distribution of the joined human-written and machine-continued texts.

4. Evaluation Metrics

We define the following evaluation metrics for each sub-task:

- For sub-task A: the main evaluation metric is Accuracy obtained by each system in the test set. Furthermore, we also report True Positive Rate (TPR) and False Positive Rate (FPR) for all systems.
- For sub-task B: the evaluation metric is the Mean Absolute Error (MAE) computed as follows:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - x_i|$$

Team	S	Affiliations	Sub-task		Runs	
			A	B	A	B
Gradient Descenders	3	University of Information Technology, Ho Chi Minh City, Vietnam; Vietnam National University, Ho Chi Minh City, Vietnam	✓	✓	3	1
Kenji-Endo	4	University of Turin, Turin, Italy; aequa-tech, Turin, Italy	✓		1	
MINDS	1	Politecnico di Torino, Turin, Italy		✓		2
Nicla	1	Independent	✓	✓	1	1
UniTor	5	University of Rome Tor Vergata, Rome, Italy	✓	✓	6	2
Stochastic Gradient Descenders	3	University of Information Technology, Ho Chi Minh City, Vietnam; Vietnam National University, Ho Chi Minh City, Vietnam	✓	✓	2	2
Pangram		Pangram Labs	✓		1	

Table 2
Team members, affiliations, participation, and number of submission for the test-phase. Column *S* reports the number of team members.

5. Participants

We received a total of 35 submissions from six unique teams. Each team could submit up to 10 runs per sub-task, with the best-performing one automatically selected as the final score. To handle submissions we use the Codabench platform [18].

Furthermore, the DeSegMa-IT task was joined by Pangram Labs[6] as a non-competing industrial participant. In Table 2, we provide an overview of the participants' affiliations, team composition, joined sub-tasks, and number of test-run submitted.

Gradient Descenders For the **sub-task A**: The team [19] employs the `UmBERTo` model[7] an encoder only language model trained on Italian data. The authors add Multi Layer Perceptron with two dense layers and tanh activation as a classification head that takes the model `[CLS]` token as input.

For the **sub-task B**: The team approaches segmentation as a token level binary-classification task. They assign to each token a human-written or machine-generated text and a train a binary classifier. At inference time, the first token in the sequence classified as human-written is selected and set as the boundary index by mapping the token boundary to its respective character. For binary classification they rely on a `DeBERTa` model fine-tuned for the Italian language.[8]

Kenji Endo For the **sub-task A**: The team [20] employs a decoder-only transformer architecture pre-trained from scratch on the Kenji-Endo dataset, exploring both a dense model and a Mixture-of-Experts (MoE) variant. The team investigates two classification paradigms for this sub-task: a discriminative fine-tuning approach, in which a classification head is trained on top of the dense model, and a generative, prompt-based approach, applicable to both the dense and MoE models. For discriminative training, the dense model is fine-tuned either for a single epoch on the full training set or for multiple epochs on a reduced subset, while the generative setting performs inference via prompting without additional fine-tuning. The dense discriminative model trained for a single epoch on the full dataset was selected for submission.

UniTor For the **sub-task A**: The system submitted by the team [21] consists of a fine-tuned version of `ModernBERT-large` trained on an augmented[9] dataset. This augmented dataset consists of 14.000 ad-

[6] https://www.pangram.com/
[7] https://github.com/musixmatchresearch/umberto
[8] https://huggingface.co/osiria/deberta-base-italian
[9] Note that this was not allowed by the competition rules, as reported in the website (https://desegma.github.io/): "Keep in mind that you should only use the training dataset we make available to train your detectors."

ditional instances: 7,000 human-written articles from the CHANGE-IT [22] corpus of Italian newspaper articles and 7,000 synthetic samples generated by translating the English RAID benchmark [23].

For the **sub-task B**: The team leverages `ModernBERT-large` fine-tuned for token-level binary classification, where each token is labeled as human-authored or machine-generated. A two-layer MLP classification head produces per-token probabilities. Rather than thresholding individual token predictions, boundary localization is performed via a change point detection procedure: the boundary token is selected by maximizing a score that aggregates log-likelihood evidence for human-authored tokens before the boundary and machine-generated tokens after it.

Nicla For the **sub-task A**: Team [24] addresses the task using `DistilBERT-base` fine-tuned for binary text classification. The model employs a standard sequence classification head, with hyperparameters selected via Bayesian optimization on a held-out validation subset to improve accuracy and generalization.

For the **sub-task B**: The team addresses the task by training a `LightGBM` regressor on sentence embeddings produced by a frozen Sentence-BERT model based on `all-MiniLM-L6-v2`, directly predicting the character index of the human–machine boundary.

Stochastic Gradient Descenders For the **sub-task A**: Team [25] frames sub-task A as a conditional single-token generation task using a decoder-only instruction-following LLM. The team fine-tunes `Qwen2.5-0.5B-Instruct` via supervised instruction tuning (SFT) with a conversation-style prompt, where the model outputs "0" for human-written or "1" for machine-generated text. Low-Rank Adaptation (LoRA) is applied to all linear layers to reduce computational cost and mitigate forgetting. At inference, predictions are obtained through greedy decoding of the first token, with a fallback to inspect raw logits if the model outputs non-numeric text.

For the **sub-task B**: The team reformulates the task as a token-level sequence labeling problem, labeling each token as human-written or machine-generated. The team fine-tunes `UmBERTo` [26] with a standard token-level classification head on top of the encoder. During inference, the model predicts per-token labels, and the boundary is extracted as the start character of the first token classified as machine-generated. Full fine-tuning is performed with standard optimization and mixed-precision training.

MINDS For the **sub-task B**: The team [27] framed the task as a token-level sequence labeling problem, using an encoder-only transformer architecture to predict human–LLM segment boundaries. They experimented with different pretrained backbones, including `BERT` and `RoBERTa` variants, under a shared training setup, finding that an Italian-specific `BERT`[10] model yielded the best performance. Token-level predictions are converted into hard labels via a post-training threshold selection step, where the decision threshold is tuned on a validation set.

Baseline For the **sub-task A**: The baseline system is based on a multilingual version of `DeBERTa-v3` [28] with a classification head. During training the backbone model is kept frozen, and only head's parameters are updated. The baseline is trained for one epoch on the whole training set.

For the **sub-task B**: The baseline system is based on a multilingual version of `DeBERTa-v3` [28] with a token-level classification head. During training the backbone model is kept frozen, and only head's parameters are updated. The baseline is trained for one epoch on the whole training set. To select the switching token, the leftmost machine-generated token is selected, and mapped to its starting character index.

[10] https://huggingface.co/dbmdz/bert-base-italian-cased

Rank	Team	**Accuracy**	TPR ↑	FPR ↓
1	Gradient Descenders	**0.9458**	0.8965	0.0048
2	Kenji-Endo	0.9426	**0.9031**	0.0176
3	UniTor*	0.9288	0.8621	0.0042
4	Nicla	0.9243	0.8598	0.0109
5	Stochastic Gradient Descenders	0.9216	0.8468	**0.0032**
	Baseline	0.8769	0.7857	0.0315
	Pangram	0.8893	0.7791	0.0000
	UniTor (late)	0.9578	0.9238	0.0081

Table 3
Team rankings and accuracies on sub-task A. Rows with a gray background indicate non-competing results. **Bold** scores indicate best results among competing participants, an underlined one indicates best results among both competing and non-competing systems.

	Sub-task A					Sub-task B				
	Encoder-only LLM	Decoder-only LLM	Italian LLM	Multilingual LLM	English-first LLM	Encoder-only LLM	Decoder-only LLM	Italian LLM	Multilingual LLM	English-first LLM
Gradient Descenders	✓		✓			✓		✓		
Kenji-Endo		✓	✓							
MINDS						✓		✓		
Nicla	✓				✓	✓				✓
UniTor	✓			✓		✓			✓	
Stochastic Gradient Descenders		✓		✓		✓		✓		
Baseline	✓			✓		✓			✓	

Table 4
Participants' design choices.

6. Results

Most teams participated in both sub-tasks, with the exception of the *Kenji-Endo* team which participated only in sub-task A and the MINDS team which only participated in sub-task B. We describe the results of the two sub-tasks separately.

6.1. Sub-task A

Table 3 reports the results obtained by participants on sub-task A and B, respectively. Note that, for the ranking of submitted systems, we consider the accuracy score. All submitted systems outperform the baseline model. The best-performing system was submitted by the team *Gradient Descenders*, which fine-tuned the Italian only LLM UmBERTo with a two-layer classification head. Notably, the second-best submission, by team *Kenji-Endo*, is achieved by the only system based on a decoder-only transformer. The decoder is pre-trained from scratch on an Italian-only corpus to assess the effectiveness of smaller LM trained on curated, language-specific data. This result underscores the potential for decoder-only models to be used for document-level tasks, such as machine-generated text detection.

The two best performing participating submissions report an accuracy marginally above 0.94 with the highest being 0.9458. This indicates that while developed systems are effective at detecting machine generated texts approximately 5% of the texts are misclassified, i.e. human-written texts are classified as machine-generated or the opposite. This is a significant limitation for a sensitive task such as detecting machine generated texts. For example an error of this kind can impact school- or work-related grading or performance reviews. Table 3 also reports a late submission from the *UniTor* team (which does

Rank	Team Name	MAE ↓
1	Stochastic Gradient Descenders	52.54
2	MINDS	56.53
3	Gradient Descenders	62.66
4	UniTor	81.60
5	Nicla	102.04
	Baseline	57.56

Table 5
Team rankings and Mean Absolute Precision scores on sub-task B. Rows with a gray background indicate non-competing results.

not count for the ranking) shows that higher accuracies are possible in our dataset, in particular they achieve an accuracy of 0.9578 supporting that novel techniques can further improve the leader-board of the DeSegMa-IT task.

To provide a deeper analysis of the kind of errors made by participating teams, it is worth analyzing models' predictions through the lens of additional metrics in addition to accuracy. For this reason, in Table 3 we report False Positive and True Positive Rates (FPR and TPR respectively).

We see that all systems have and FPR lower than 2% with the highest value 0.178 shown by *Kenji-Endo* and the lowest by *Stochastic Gradient Descenders* 0.0033. Based on these results, in the best case scenario every 1,000 human texts, 3 would be wrongly attributed to generative AI, while in the worst case this would happen for about 18 human-written texts. These results highlight that participating systems occasionally attribute human-written texts to AI systems, this is not desirable as it undermines the authorship of "authentic" content created by humans.

We also investigate which participants' choices resulted in better performance. Specifically, Table 4 reports the main decisions made by each team when developing their system. As expected all teams used language models to tackle DeSegMa-IT, therefore we report which teams used encoder-only and/or decoder-only language models and the language each model is trained on. For sub-task A we see an almost even split between encoder- and decoder-only language models (3 encoder and 2 decoder) and interestingly the first and second best performing results are obtained by teams using an encoder-only and a decoder-only language model respectively. Concerning the model language, we see a clear benefit from using models that have been trained on Italian texts, a choice made by the two best-performing submissions.

6.1.1. Pangram Labs' Participation

Pangram[11] is an online tool for the automatic detection of Machine generated texts. They adopt a transformer-based classifier trained on a dataset they develop to detect both closed source Generative AI systems such as `GPT-4-0613` and open-source ones such as `Llama-2-70b-chat` [29]. Their approach to the DeSegMa-IT task is different from others because their system does not have access to the training set which was released at a time when the Pangram detector was already available. Due to this difference, their results are reported in Table 3 as non-competing. They have lower accuracy than models trained on the training set developed specifically for the DeSegMa-IT task, however they are the only system with 0 FPR. This result indicates that their system never attributes human-written texts to Generative AI.

6.2. Sub-task B

Table 5 reports the results for Sub-task B. Two out of the five submitted systems outperform the baseline. The best performing system was submitted by the *Stochastic Gradient Descenders* team, the team fine-tuned the Italian only LLM UmBERTo and adopted a leftmost machine-generated token decision strategy. The second-best system, by team *MINDS*, also relies on an Italian-specific variant of BERT, confirming

[11] https://www.pangram.com/

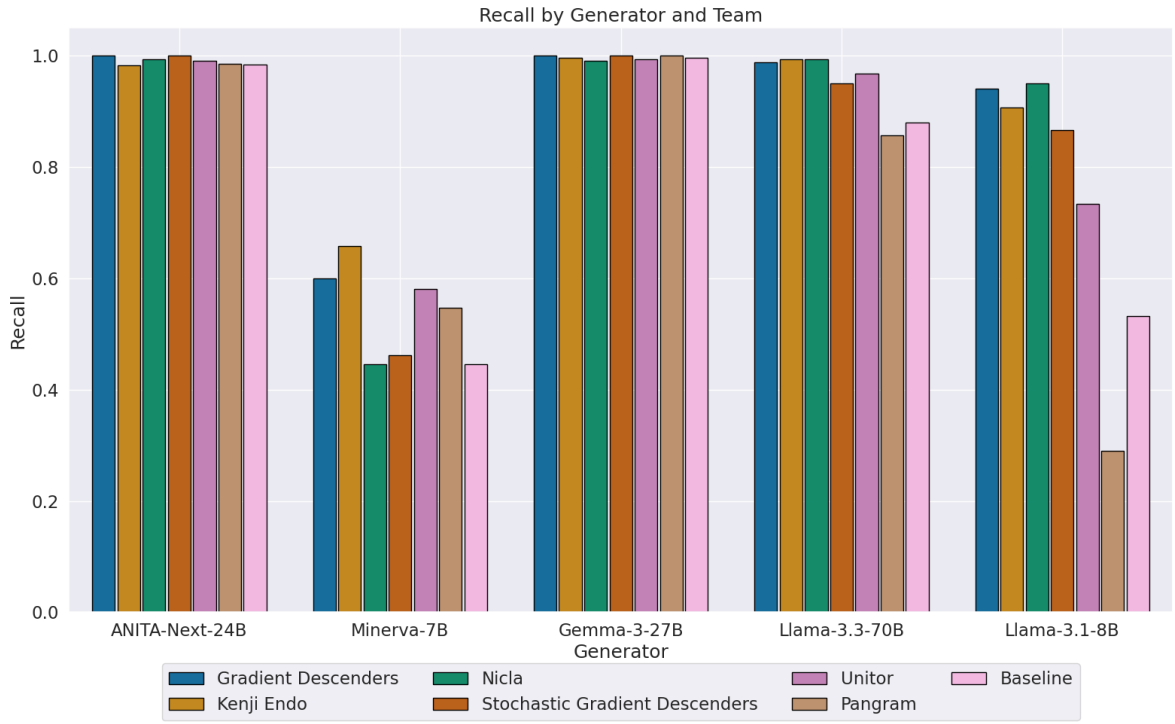


Figure 7: Sub-task A: Per-generator recall for each participant.

the relevance of language-specific pre-training for fine-grained tasks such as the human-MGT text segmentation. Among the remaining teams, the regression-based approach proposed by team *Nicla* does not achieve competitive performance, highlighting the effectiveness of framing the segmentation task as a sequence-labeling problem.

Also for sub-task B Table 4 reports the design choices, in this case there is a clear preference for encoder-only language models. Same as for sub-task A there is and advantage in using Italian-first models, the choice made by the two best-performing teams.

7. Discussion

The task of detecting machine-generated texts is a challenging tasks mostly due to the general purpose abilities of modern LLMs and because of the large number of available models. DeSegMa-IT's sub-task A challenged participants to train detectors that are able to address specifically this difficulty by including synthetic texts generated by different systems. In particular, while train and test set do not share any data, they share some of the LLMs used to generate synthetic texts. As a result, the accuracy of the detectors is high but not close to a perfect score. Highlighting that sensitivity to the LLM used to generate texts is a key factor to account for when training MGT detectors.

To better understand how different generators affect detection performance, we analyze systems separately for each generator, as reported in Figure 7. This analysis is conducted exclusively on the subset of machine-generated texts, with human-authored texts excluded. As no negative instances are present in this setting, we report recall (i.e., the true positive rate) for each generator. This analysis reveals interesting patterns. All systems achieve high recall on texts generated by models based on English-first pretrained models: on texts generated by `Llama-3.3-70B`, and `Gemma-3-27B` the average recall exceeds 90%. Similar results are observed for English pretrained models later fine-tuned on Italian, `ANITA-Next-24B` is easy to detect, with systems maintaining a recall above 90%. In contrast, models pre-trained from scratch on Italian data are more challenging to detect, resulting in lower recall scores, close to 50%. The only model that deviates from this pattern is `Llama-3.1-8B` which, despite being

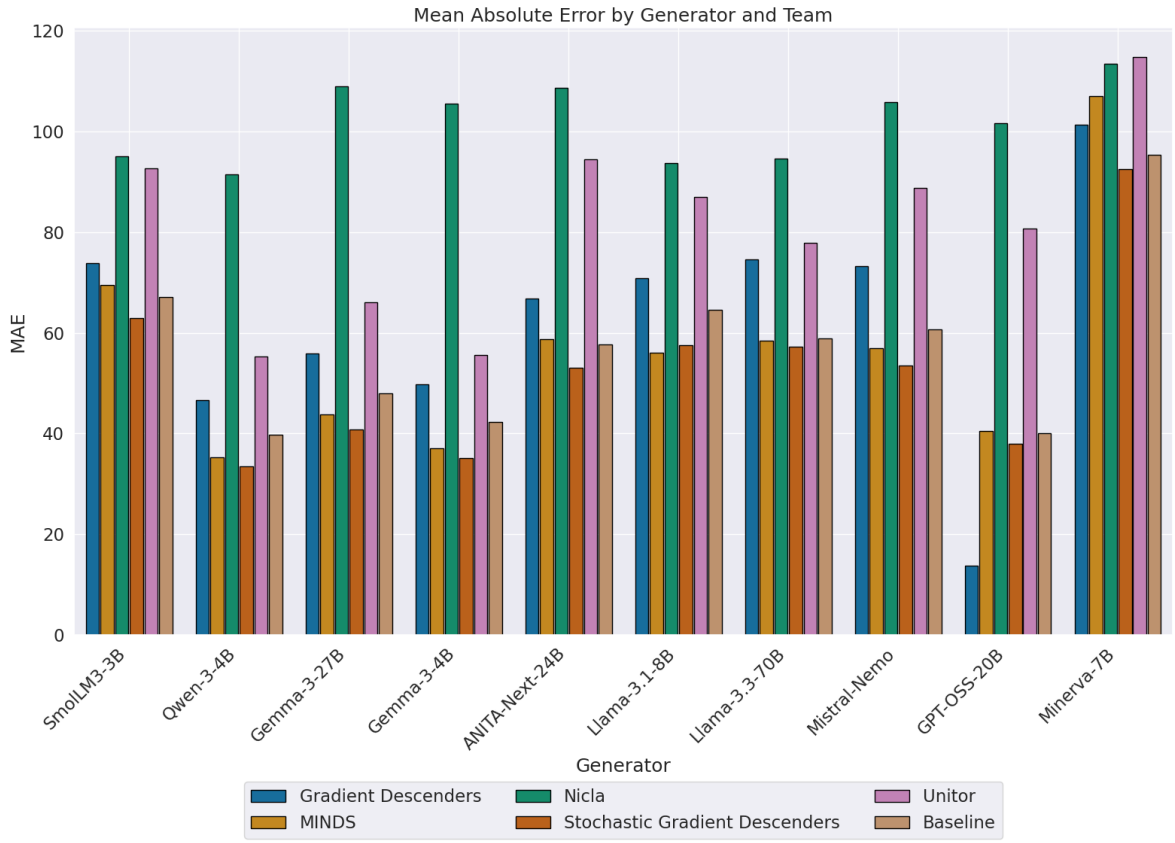


Figure 8: Sub-task B: Per-generator MAE for each participant.

pretrained on English, evades some of the detectors. We interpret this as evidence of unexpected generation patterns when the model produces Italian texts.

When performing MGT detection there are two possible errors, human-written texts classified as AI-generated and the opposite, AI-generated texts classified as human-written. In real applications the first type of error can be more harmful, since it results in genuine human effort being misattributed to AI systems. To quantify how often this happens we compute the FPR of participating systems and we find that all systems are between 0.3% and 2% FPR. While this shows that the proposed systems rarely make this type of error there are instances of human-written texts attributed to AI.

The only system showing 0 FPR is the Pangram detector which is however less accurate than others, this can be due to not relying on the DeSegMa-IT training set. Restricted to our test set, Pangram never attributes human-written texts to AI but it is more prone to attributing MGT texts to humans. This is a desirable design choice necessary for responsible deployment of MGT detectors.

Sub-task B, segmenting human-written and machine-generated texts, requires participants to identify the character where a text that is human-written in its first part switches to machine-generated. For this task, we measure performance through character-level mean absolute error, which corresponds to the number of offset characters between the predicted switch and the real one. We see that achieving an error inferior to 50 characters has proven challenging for participants, with the best system achieving a MAE of 52.54. Unlike sub-task A, identifying which models are harder to segment is not evident. Figure 8 shows the MAE achieved by each participant submission restricted to single models. There are fewer clear patterns compared to sub-task A, the main observations we draw is that `SmolLM3-3B` and `Minerva 7B` are more difficult to segment, with an average MAE above 70 characters, while `GPT-OSS-20B` is relatively easier with average MAE close to 40. On the remaining models, the average MAE of participants is evenly spread between 40 and 70, showing that the text generated by all models

is comparably challenging to segment from human-written text.

Acknowledgments

Giovanni Puccetti is fully funded by the Italian Ministry of University and Research under the PNRR project ITSERR (CUP B53C22001770006). Andrea Pedrotti is fully funded by the European Union - NextGenerationEU through PNRR (CUP B53C22001760006) "SoBigData.it: Strengthening the Italian RI for Social Mining and Big Data Analytics" (SoBigData.it).

Declaration on Generative AI

During the preparation of this work, the authors used GPT-4 in order to: Grammar and spelling check; Improve writing style.

References

[1] OpenAI, GPT-4 Technical Report, 2023. URL: https://arxiv.org/abs/2303.08774.

[2] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan, et al., The llama 3 herd of models, arXiv preprint arXiv:2407.21783 (2024).

[3] DeepSeek-AI, Deepseek-v3 technical report, 2024. URL: https://arxiv.org/abs/2412.19437. arXiv:2412.19437.

[4] L. Dugan, D. Ippolito, A. Kirubarajan, S. Shi, C. Callison-Burch, Real or fake text?: Investigating human ability to detect boundaries between human-written and machine-generated text, Proceedings of the AAAI Conference on Artificial Intelligence 37 (2023) 12763–12771. URL: https://ojs.aaai.org/index.php/AAAI/article/view/26501. doi:10.1609/aaai.v37i11.26501.

[5] G. Puccetti, A. Rogers, C. Alzetta, F. Dell'Orletta, A. Esuli, AI 'news' content farms are easy to make and hard to detect: A case study in italian, in: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, 2024, p. 15312–15338. URL: http://dx.doi.org/10.18653/v1/2024.acl-long.817. doi:10.18653/v1/2024.acl-long.817.

[6] E. Kasneci, K. Sessler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, U. Gasser, G. Groh, S. Günnemann, E. Hüllermeier, S. Krusche, G. Kutyniok, T. Michaeli, C. Nerdel, J. Pfeffer, O. Poquet, M. Sailer, A. Schmidt, T. Seidel, M. Stadler, J. Weller, J. Kuhn, G. Kasneci, Chatgpt for good? on opportunities and challenges of large language models for education, Learning and Individual Differences 103 (2023) 102274. URL: https://www.sciencedirect.com/science/article/pii/S1041608023000195. doi:https://doi.org/10.1016/j.lindif.2023.102274.

[7] M. Corazza, G. Longo, L. Zilli, E. Di Sante, S. Sapienza, M. Palmirani, Hybrid ai enhancing european drafting legislation for a better regulation, in: 2025 Eleventh International Conference on eDemocracy & eGovernment (ICEDEG), IEEE, 2025, pp. 106–113.

[8] Y. Li, Q. Li, L. Cui, W. Bi, Z. Wang, L. Wang, L. Yang, S. Shi, Y. Zhang, MAGE: Machine-generated text detection in the wild, in: L.-W. Ku, A. Martins, V. Srikumar (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 36–53. URL: https://aclanthology.org/2024.acl-long.3/. doi:10.18653/v1/2024.acl-long.3.

[9] X. Hu, P.-Y. Chen, T.-Y. Ho, RADAR: robust ai-text detection via adversarial learning, in: Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023, 2023.

[10] M. Abassy, K. Elozeiri, A. Aziz, M. N. Ta, R. V. Tomar, B. Adhikari, S. E. D. Ahmed, Y. Wang, O. Mohammed Afzal, Z. Xie, J. Mansurov, E. Artemova, V. Mikhailov, R. Xing, J. Geng, H. Iqbal, Z. M. Mujahid, T. Mahmoud, A. Tsvigun, A. F. Aji, A. Shelmanov, N. Habash, I. Gurevych, P. Nakov, LLM-DetectAIve: a tool for fine-grained machine-generated text detection, in: D. I. Hernandez Farias,

T. Hope, M. Li (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 336–343. URL: https://aclanthology.org/2024.emnlp-demo.35/. doi:10.18653/v1/2024.emnlp-demo.35.

[11] A. Pedrotti, M. Papucci, C. Ciaccio, A. Miaschi, G. Puccetti, F. Dell'Orletta, A. Esuli, Stress-testing machine generated text detection: Shifting language models writing style to fool detectors, in: W. Che, J. Nabende, E. Shutova, M. T. Pilehvar (Eds.), Findings of the Association for Computational Linguistics: ACL 2025, Association for Computational Linguistics, Vienna, Austria, 2025, pp. 3010–3031. URL: https://aclanthology.org/2025.findings-acl.156/. doi:10.18653/v1/2025.findings-acl.156.

[12] Y. Wang, J. Mansurov, P. Ivanov, J. Su, A. Shelmanov, A. Tsvigun, O. M. Afzal, T. Mahmoud, G. Puccetti, T. Arnold, Semeval-2024 task 8: Multidomain, multimodel and multilingual machine-generated text detection, in: A. K. Ojha, A. S. Dogruöz, H. T. Madabushi, G. D. S. Martino, S. Rosenthal, A. Rosá (Eds.), Proceedings of the 18th International Workshop on Semantic Evaluation, SemEval@NAACL 2024, Mexico City, Mexico, June 20-21, 2024, Association for Computational Linguistics, 2024, pp. 2057–2079. URL: https://doi.org/10.18653/v1/2024.semeval-1.279. doi:10.18653/V1/2024.SEMEVAL-1.279.

[13] Y. Wang, A. Shelmanov, J. Mansurov, A. Tsvigun, V. Mikhailov, R. Xing, Z. Xie, J. Geng, G. Puccetti, E. Artemova, J. Su, M. N. Ta, M. Abassy, K. A. Elozeiri, S. E. D. A. El Etter, M. Goloburda, T. Mahmoud, R. V. Tomar, N. Laiyk, O. Mohammed Afzal, R. Koike, M. Kaneko, A. F. Aji, N. Habash, I. Gurevych, P. Nakov, GenAI content detection task 1: English and multilingual machine-generated text detection: AI vs. human, in: F. Alam, P. Nakov, N. Habash, I. Gurevych, S. Chowdhury, A. Shelmanov, Y. Wang, E. Artemova, M. Kutlu, G. Mikros (Eds.), Proceedings of the 1stWorkshop on GenAI Content Detection (GenAIDetect), International Conference on Computational Linguistics, Abu Dhabi, UAE, 2025, pp. 244–261. URL: https://aclanthology.org/2025.genaidetect-1.27/.

[14] Y. Wang, J. Mansurov, P. Ivanov, J. Su, A. Shelmanov, A. Tsvigun, O. M. Afzal, T. Mahmoud, G. Puccetti, T. Arnold, A. F. Aji, N. Habash, I. Gurevych, P. Nakov, M4gt-bench: Evaluation benchmark for black-box machine-generated text detection, 2024. arXiv:2402.11175.

[15] F. Cutugno, A. Miaschi, A. P. Aprosio, G. Rambelli, L. Siciliani, M. A. Stranisci, Evalita 2026: Overview of the 9th evaluation campaign of natural language processing and speech tools for italian, in: Proceedings of the Ninth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2026), CEUR.org, Bari, Italy, 2026.

[16] E. Mitchell, Y. Lee, A. Khazatsky, C. D. Manning, C. Finn, Detectgpt: zero-shot machine-generated text detection using probability curvature, in: Proceedings of the 40th International Conference on Machine Learning, ICML'23, JMLR.org, 2023.

[17] L. D. Mattei, M. Cafagna, F. Dell'Orletta, M. Nissim, A. Gatt, CHANGE-IT @ EVALITA 2020: Change headlines, adapt news, generate, in: V. Basile, D. Croce, M. D. Maro, L. C. Passaro (Eds.), Proceedings of Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020), CEUR.org, Online, 2020.

[18] Z. Xu, S. Escalera, A. Pavão, M. Richard, W.-W. Tu, Q. Yao, H. Zhao, I. Guyon, Codabench: Flexible, easy-to-use, and reproducible meta-benchmark platform, Patterns 3 (2022) 100543. URL: https://www.sciencedirect.com/science/article/pii/S2666389922001465. doi:https://doi.org/10.1016/j.patter.2022.100543.

[19] T. P. T. Nhan, B. H. Son1, D. V. Thin, Gradient descenders at DeSegMa-IT: Leveraging monolingual transformer for LLM-generated text detection and boundary identification, in: Proceedings of the Ninth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2026), CEUR.org, Bari, Italy, 2026.

[20] C. J. Scozzaro, M. Rinaldi, G. Mittone, M. A. Stranisci, Kenji-Endo: a BabyLM at EVALITA 2026, in: Proceedings of the Ninth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2026), CEUR.org, Bari, Italy, 2026.

[21] F. Borazio, G. D. Luca, D. Pasquini, D. Croce, R. Basili, UniTor at DeSegMa-IT Analyzing supervision and encoder representations for italian machine-generated text detection, in: Proceedings

of the Ninth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2026), CEUR.org, Bari, Italy, 2026.
[22] L. D. Mattei, M. Cafagna, F. Dell'Orletta, M. Nissim, A. Gatt, CHANGE-IT @ EVALITA 2020: Change headlines, adapt news, generate (short paper), in: V. Basile, D. Croce, M. D. Maro, L. C. Passaro (Eds.), Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020), Online event, December 17th, 2020, volume 2765 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020. URL: https://ceur-ws.org/Vol-2765/paper169.pdf.
[23] K. Dubey, Evaluating the fairness of task-adaptive pretraining on unlabeled test data before few-shot text classification, in: D. Hupkes, V. Dankers, K. Batsuren, A. Kazemnejad, C. Christodoulopoulos, M. Giulianelli, R. Cotterell (Eds.), Proceedings of the 2nd GenBench Workshop on Generalisation (Benchmarking) in NLP, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 1–26. URL: https://aclanthology.org/2024.genbench-1.1/. doi:10.18653/v1/2024.genbench-1.1.
[24] N. Auletta, Nicla at DeSegMa-IT: DistilBERT for MGT detection and LightGBM regressor for HMT segmentation., in: Proceedings of the Ninth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2026), CEUR.org, Bari, Italy, 2026.
[25] H. N. Phu, B. H. Son, D. V. Thin, Stochastic Gradient Descenders at DeSegMa-IT: Instruction-tuned LLM and token classification for MGT detection and boundary localization, in: Proceedings of the Ninth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2026), CEUR.org, Bari, Italy, 2026.
[26] L. Parisi, S. Francia, P. Magnani, UmBERTo: an italian language model trained with whole word masking, https://github.com/musixmatchresearch/umberto, 2020.
[27] F. Giobergia, MINDS at DeSegMa-IT: Detecting human-LLM authorship switches via token-level classification, in: Proceedings of the Ninth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2026), CEUR.org, Bari, Italy, 2026.
[28] P. He, J. Gao, W. Chen, DeBERTav3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing, in: The Eleventh International Conference on Learning Representations, 2023. URL: https://openreview.net/forum?id=sE7-XhLxHA.
[29] B. Emi, M. Spero, Technical report on the pangram ai-generated text classifier, 2024. URL: https://arxiv.org/abs/2402.14873. arXiv:2402.14873.

A. Generative Models Details

Table 6 reports the four models used to create the machine-generated part of the training set of Sub-task A. Table 7 reports the four models used to create the machine-generated part of the test set of Sub-task A.

A.1. Sub-task A

Model	Hugging Face URL
`Qwen3-4B-Instruct`	Qwen/Qwen3-4B-Instruct-2507
`Gemma-3-4B-it`	google/gemma-3-4b-it
`Nemo-Instruct`	mistralai/Mistral-Nemo-Instruct-2407
`Gpt-oss-20B`	openai/gpt-oss-20b

Table 6
Training models used, with their relative Hugging Face URLs.

Model	Relative URL
ANITA-NEXT-24B	m-polignano/ANITA-NEXT-24B-Magistral-2506-ITA
Llama-3.3-70B-Instruct	meta-llama/Llama-3.3-70B-Instruct
Minerva-7B	sapienzanlp/Minerva-7B-instruct-v1.0
Gemma-3-27B-it	google/gemma-3-27b-it

Table 7
Models selected for evaluation, with Hugging Face relative URLs.

A.2. Sub-task B

Table 8 reports the nine models used to generate the continuation of human-written news in the train and test sets of Sub-task B.

Model	Relative URL
SmolLM3-3B	HuggingFaceTB/SmolLM3-3B
Qwen3-4B-Instruct	Qwen/Qwen3-4B-Instruct-2507
Gemma-3-27b-it	google/gemma-3-27b-it
ANITA-NEXT-24B	m-polignano/ANITA-NEXT-24B-Magistral-2506-ITA
Llama-3.1-8B-Instruct	meta-llama/Llama-3.1-8B-Instruct
Llama-3.3-70B-Instruct	meta-llama/Llama-3.3-70B-Instruct
Nemo-Instruct	mistralai/Mistral-Nemo-Instruct-2407
Gpt-oss-20b	openai/gpt-oss-20b
Minerva-7B	sapienzanlp/Minerva-7B-instruct-v1.0

Table 8
Generator models used in sub-task B, with Hugging Face relative URLs.