

# UniTor at DeSegMa-It: Analyzing Supervision and Encoder Representations for Italian Machine-Generated Text Detection

Federico Borazio<sup>\*1</sup>, Giacomo De Luca<sup>\*1</sup>, Daniele Pasquini<sup>\*1</sup>, Danilo Croce<sup>1</sup> and Roberto Basili<sup>1</sup>

<sup>1</sup>Department of Enterprise Engineering, University of Rome Tor Vergata  
Via del Politecnico 1, 00133, Rome, Italy

<sup>\*</sup>Equal contribution.

## Abstract

This paper describes the UniTor systems submitted to the DeSegMa-IT shared task at EVALITA 2026. Our study investigates the impact of unsupervised versus supervised approaches for machine-generated text (MGT) detection and segmentation in Italian, and analyzes which components of encoder-only architectures are most relevant for these tasks. For Subtask A, we compare a data-agnostic, zero-shot detector based on Fast-DetectGPT with supervised ModernBERT-based classifiers fine-tuned on Italian data. For Subtask B, we address boundary localization via token-level classification followed by change point detection, evaluating both supervised transformer-based models and an unsupervised token-level adaptation of Fast-DetectGPT. While our submitted systems achieve mid-range leaderboard performance, post-submission experiments reveal a key divergence: for document-level detection, model size proves decisive, with the English-centric ModernBERT-large outperforming the MultilingualModernBERT-base alternative; for boundary localization, multilingual pretraining yields substantial gains, reducing error by over 25%. These results explore the tradeoffs between robustness and task-specific supervision, and show that document-level detection relies on structural features that generalize across languages, whereas token-level segmentation requires language-specific models.

## Keywords

Machine Text Detection, Encoder Models, Layer Probing, Fine-Tuning, Zero-Shot Detection

## 1. Introduction

The rapid evolution of Large Language Models (LLMs) has democratized the ability to generate text that is increasingly difficult to distinguish from human writing. While this technological progress enables a wide range of beneficial applications, it has also substantially lowered the cost and expertise required to produce large volumes of synthetic content, facilitating its use in deceptive, manipulative, or low-quality information campaigns. Recent state-of-the-art open-weight models, such as DeepSeek [1] and Kimi [2], provide high-quality text generation at scale, enabling the automatic production of fluent and coherent narratives across a wide variety of domains. As a consequence, machine-generated text (MGT) is no longer confined to clearly artificial or low-effort outputs, but is often embedded seamlessly within otherwise legitimate content streams. This phenomenon has attracted increasing public attention, as reflected by the designation of “slop” as the Word of the Year for 2025 by the Merriam-Webster dictionary<sup>1</sup>.

At the same time, recent advances in model alignment manipulation have further complicated the detection landscape. Uncensored or weakly aligned variants of open-source models are increasingly accessible, either through architectural interventions (e.g., ablation of refusal mechanisms) or post-training fine-tuning procedures. Prior work has shown that refusal behavior in aligned LLMs can

---

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

✉ borazio@ing.uniroma2.it (F. Borazio<sup>\*</sup>); deluca@ing.uniroma2.it (G. De Luca<sup>\*</sup>); daniele.pasquini@uniroma2.eu (D. Pasquini<sup>\*</sup>); croce@info.uniroma2.it (D. Croce); basili@info.uniroma2.it (R. Basili)

🆔 0009-0000-0193-2131 (F. Borazio<sup>\*</sup>); 0009-0004-5469-3364 (G. De Luca<sup>\*</sup>); 0009-0006-8172-3228 (D. Pasquini<sup>\*</sup>); 0000-0001-9111-1950 (D. Croce); 0000-0001-5140-0694 (R. Basili)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

<sup>1</sup><https://www.merriam-webster.com/wordplay/word-of-the-year>

be mediated by low-dimensional subspaces in the residual stream and can therefore be substantially altered without significantly degrading generation quality [3]. As a result, machine-generated texts can easily include stylistic features, such as profanity, idiosyncratic phrasing, or unconventional discourse patterns, that were previously strong cues of human authorship.

These developments pose a critical challenge for existing detection systems. Detectors trained primarily on RLHF-aligned models and conventional prompting strategies often exhibit limited robustness to paraphrasing [4], adversarial rewriting [5], or stylistic deviations introduced by weakly aligned generators. Consequently, reliable MGT detection must operate under realistic, non-I.I.D. conditions in which both the generating models and their alignment properties are unknown.

While the proliferation of synthetic content is a global phenomenon, defensive mechanisms are hindered by a severe asymmetry in available linguistic resources. A central bottleneck lies in the overwhelming predominance of English in both detection benchmarks and training corpora. Foundational datasets for MGT detection, such as TuringBench [6] and the HC3 corpus [7], are almost exclusively English-centric, reflecting a broader trend in large-scale language modeling.

Although recent evaluation campaigns have introduced multilingual components [8, 9], languages other than English remain substantially underrepresented. For Italian in particular, detection strategies often rely on translated data or small ad-hoc datasets, which fail to capture language-specific syntactic, morphological, and discourse-level artifacts [10]. This resource gap creates a critical blind spot: without native benchmarks that simulate adversarial “in the wild” conditions, the robustness and reliability of state-of-the-art MGT detectors for Italian remain largely unverified.

To address these challenges within the Italian landscape, this study targets the **DeSegMa-IT** [11] shared task in Evalita 2026 [12]. DeSegMa-IT is designed to benchmark the robustness of systems against realistic, non-I.I.D. scenarios and is structured into two main sub-tasks:

- **Subtask A: MGT Detection in the Wild.** This subtask is formulated as a *binary document-level classification* problem, where each input text must be labeled as either human-written or machine-generated. Given a complete document, the system is required to assign a single label to the entire text, as illustrated in Fig. 1. To explicitly simulate “in-the-wild” conditions, the evaluation setting departs from standard i.i.d. assumptions. Test documents are sampled from semantic domains that differ from those observed during training and are generated by *undisclosed* large language models. These generators range from vanilla pre-trained models to variants fine-tuned (e.g., via DPO) to better mimic human linguistic distributions. As a result, human- and machine-written documents often discuss highly similar topics, forcing detection systems to rely on subtle stylistic and discourse-level cues rather than topical differences.
- **Subtask B: Human–Machine Text Segmentation.** This subtask is formulated as a *fine-grained boundary localization* problem. Given a document containing both human-written and machine-generated text, the system must identify the exact *character index* at which the machine-generated continuation begins. Each input consists of a variable-length human-authored prompt followed by a model-generated completion. Unlike Subtask A, which requires a single document-level label, this setting demands precise localization within the text, simulating real-world scenarios where MGT is subtly embedded into otherwise authentic content. Fig. 2 shows a representative instance, illustrating how the transition between human and machine text can occur seamlessly within an otherwise coherent document.

A central open question in machine-generated text detection concerns the trade-off between *robustness* and *adaptation*. While supervised models can be fine-tuned to specific languages and domains, their performance may degrade under distribution shift, whereas data-agnostic methods promise greater stability at the cost of accuracy. Understanding this trade-off is particularly critical for under-resourced languages such as Italian, where large-scale native benchmarks are scarce. To investigate this question, we submitted both supervised and unsupervised systems to the DeSegMa-IT shared task, enabling a controlled comparison under identical evaluation conditions.

For Subtask A (document-level detection), we evaluated a zero-shot, domain-agnostic detector based on Fast-DetectGPT [13], a probabilistic method derived from DetectGPT [14] that exploits differences

### Example: Subtask A - Machine Generated Text Detection in the Wild

#### Sample 1: Human-Written (Label 0)

*“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. [...]”*

#### Sample 2: Machine-Generated (Label 1)

*“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”. [...]”*

**Goal:** The model must classify the entire document binary (0 vs 1). Note the subtle topical similarities making the task challenging.

**Figure 1:** Instances from the DeSegMa-IT dataset (Subtask A). Both texts discuss similar political topics, requiring the model to detect subtle stylistic or semantic incoherencies rather than simple topic shifts.

### Example: Subtask B - Human-Machine Segmentation

#### Input Document (Mixed Source)

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...

- **Human Prompt (Black):** Ends at “...ha fissato ”
- **Machine Continuation (Red):** Starts at “al 10 agosto...”
- **Target Label:** Character Index 103

**Figure 2:** Instance from Subtask B. The system receives the full text and must predict the character index (103) where the machine-generated continuation begins.

in token-level likelihood distributions between human and machine-generated text. In parallel, we developed supervised classifiers by fine-tuning ModernBERT [15], an English-centric encoder optimized for long-context processing, to assess how effectively cross-lingual transfer can bridge the Italian language gap. For Subtask B (human-machine segmentation), we framed boundary localization as a token-level binary classification problem followed by a change point detection procedure. We compared supervised encoders fine-tuned on the task against an unsupervised, token-level adaptation of Fast-DetectGPT, extending its document-level scoring mechanism to identify transitions within mixed-source texts.

Using the DeSegMa-IT benchmark as a controlled testbed, this work investigates the design space of machine-generated text detection in Italian under realistic, non-I.I.D. conditions. Our study focuses on three main aspects. First, we provide a systematic comparison between data-agnostic, zero-shot detection methods and supervised encoder-based models, analyzing their robustness when both the generating model and the data distribution are unknown. Second, we assess the effectiveness of supervised cross-lingual transfer by fine-tuning English-centric encoders on Italian data and comparing them against unsupervised baselines. Third, we analyze which components of encoder-only architectures contribute most to document-level detection performance through controlled layer-wise ablation experiments.

Overall, our experiments highlight a consistent hierarchy across both tasks: supervised encoder-based models substantially outperform the data-agnostic Fast-DetectGPT baseline in both document-level

detection (Subtask A) and fine-grained segmentation (Subtask B). While supervision proves superior for Italian MGT in the wild, we identify input length as a critical boundary condition, with short texts accounting for a large portion of residual errors. Crucially, through controlled layer-wise ablations, we show that full model depth is not a prerequisite for effectiveness: intermediate encoder layers already encode task-relevant signals that enable competitive detection performance even without relying on final-layer representations.

The remainder of this paper is organized as follows. Section 2 reviews related work, situating our study within the broader landscape of MGT detection. Section 3 describes the submitted systems, detailing both the unsupervised and supervised methodologies. Section 4 presents the experimental results and discussion, including the layer-wise ablation analysis. Finally, Section 5 concludes the paper and outlines directions for future research.

## 2. Related Work

Approaches to machine-generated text (MGT) detection are commonly divided into *supervised* classifiers and *zero-shot* (data-agnostic) methods [13]. Supervised detectors, typically trained on labeled corpora of human and synthetic text, often achieve strong in-domain performance [16], but are known to generalize poorly under distribution shift, particularly when confronted with unseen generators, domains, or decoding strategies [17]. This limitation is especially problematic in real-world scenarios, where the generating model and its alignment properties are unknown [18].

To address robustness, recent benchmarks have emphasized scale and adversarial variation. The RAID dataset [19], comprising over 10 million documents, introduces a wide range of text-level attacks such as homoglyph substitution, whitespace manipulation, and translation-based perturbations. However, a key limitation of such resources is that their synthetic texts are mostly generated using models that lag behind current state-of-the-art open-weight generators (e.g., GPT-2, GPT-3, LLaMA 2, Mistral 7B). As a result, detectors trained on these benchmarks may overfit artifacts that are no longer representative of modern, inexpensive, and highly capable models, raising concerns about the ecological validity of “in-the-wild” settings.

An alternative line of work focuses on *generator-agnostic* signals derived from language model probabilities. DetectGPT [14] exploits the observation that machine-generated text tends to occupy local maxima in the likelihood landscape of models trained on similar data, distinguishing it from perturbed variants. Fast-DetectGPT [13] extends this idea by introducing *conditional probability curvature*, showing that machine-generated text becomes increasingly predictable as generation progresses, while human-authored text remains comparatively flat. These probabilistic approaches exhibit improved robustness across generators, but their sensitivity to decoding parameters, short contexts, increase in temperature parameter and adversarial rewriting remains an open challenge.

Beyond probabilistic cues, recent work highlights *syntactic structure* as a potentially more stable discriminator between human and machine-generated text. Using *Universal Dependencies*, Guo et al. [20] show that LLM-generated text exhibits systematic syntactic divergences from human writing across languages, often retaining an “English accent” even when generating non-English text. For Italian in particular, prior studies report that model performance degrades with increasing syntactic complexity, such as inversions or post-verbal subjects [21]. Given Italian’s rich morphology and relatively flexible word order, such deviations may provide discriminative signals that are less sensitive to surface-level paraphrasing. These observations suggest that intermediate syntactic representations, rather than surface fluency alone, may offer more stable cues for machine-generated text detection under paraphrasing or stylistic manipulation.

These issues are amplified in lower-resourced languages, where both training data and evaluation benchmarks are scarce [22, 23]. For Italian, the only publicly documented evaluation of MGT detection to date is the multilingual setting of Task 8 at SemEval 2024 [8], where Italian appeared as a *surprise language*. That evaluation relied on a relatively small dataset (approximately 3,000 human and 3,000 machine-generated texts), with synthetic data produced by a fine-tuned LLaMA 2 70B model [24],

and did not explicitly model adversarial conditions or generator diversity. No large-scale adversarial benchmark comparable to RAID currently exists for Italian, and most available resources focus on general NLP evaluation rather than robustness in machine-text detection [25, 26]. As a result, the current empirical understanding of MGT detection for Italian remains limited to small-scale, non-adversarial settings.

In this context, data-agnostic approaches are particularly attractive. Prior work suggests that some probabilistic detection signals transfer across languages: DetectGPT scores have been shown to correlate across translations of the same text, including Italian [27, 14]. At the same time, the evidence on syntactic divergence suggests that language-specific structural cues may be crucial for reliable detection in morphologically rich languages.

Building on these insights, our work combines zero-shot probabilistic detectors with supervised encoder-based models and layer-wise analysis. We focus on ModernBERT [15], whose alternation of local and global attention layers is well-suited to capturing both document-level coherence and intermediate syntactic representations. By jointly evaluating data-agnostic and supervised approaches on the same benchmark and analyzing which internal representations drive performance, we aim to clarify not only *whether* MGT detection works for Italian, but *which signals* are effective, *where they emerge within encoder representations*, and *under which conditions* they remain reliable.

### 3. Description of the Systems

This section describes the systems submitted to the DeSegMa-IT shared task. We outline the modeling assumptions and architectural choices adopted for each subtask, contrasting data-agnostic and supervised approaches under identical evaluation conditions. Beyond system description, this section is structured to highlight how specific design factors, such as context granularity, representation depth, and training data diversity, shape robustness and performance in machine-generated text detection.

#### 3.1. Subtask A: Human–Machine Text Detection

Subtask A requires binary document-level classification of human-written versus machine-generated text under realistic, non-I.I.D. conditions. Rather than assuming access to the generating model or domain, the task challenges systems to operate under a distribution shift. Accordingly, we approach detection from two complementary perspectives: a data-agnostic, zero-shot baseline and supervised encoder-based models trained to adapt to Italian data.

##### 3.1.1. Unsupervised Setting: Fast-DetectGPT

For Subtask A, we use Fast-DetectGPT [13] as a data-agnostic baseline. Fast-DetectGPT is a zero-shot detector that relies on token-level probability estimates produced by a language model to distinguish human-authored from machine-generated text.

The core idea is that machine-generated text tends to follow the probability distribution of any language model more consistently than human text. Given a text  $x = (x_1, \dots, x_T)$  and a scoring model  $p_\theta$ , the method analyzes each token in context. At position  $t$ , it compares the log-probability of the observed token,  $\log p_\theta(x_t \mid x_{<t})$ , with the expected log-probability under the same predictive distribution, corresponding to the negative entropy of  $p_\theta(\cdot \mid x_{<t})$ . This yields a local discrepancy score indicating whether the token is more predictable than an average continuation according to the model.

Since this computation is performed at the token level, Fast-DetectGPT produces a sequence of discrepancy scores. To obtain a document-level decision, these scores are aggregated across the sequence and normalized for length and variance, yielding a single scalar score  $d(x)$ . Texts that consistently align with the model’s expectations produce positive normalized scores, while human-authored texts tend to fluctuate around zero. Final classification is obtained by thresholding  $d(x)$ .

Token probabilities, and thus the detection signal, depend on the choice of the scoring language model. Using a model close to the generator strengthens this signal. Accordingly, Fast-DetectGPT



distinguishes between a *white-box* setting, where the scoring model matches or closely approximates the generator, and a *black-box* setting, where a surrogate model is used. Here, *white-box* does not imply access to model parameters, but rather membership in the same model family.

In our experiments, we evaluated multiple scoring models (Llama-3.1-8B-Instruct, Gemma-3-12B-IT, Falcon-7B-Instruct, BLOOM-7B1) on the training split. Llama-3.1-8B-Instruct achieved the highest accuracy (0.948) and was selected for the final submission as a quasi-white-box configuration, since the generators in the test set are undisclosed.

### 3.1.2. Supervised Setting: Encoder-Based Classification

Since Subtask A is formulated as a document-level binary classification problem, we adopt a supervised encoder-based approach. Transformer encoders are a natural choice in this setting: self-attention enables the integration of evidence across long contexts, while internal representations can encode syntactic and structural regularities that differ between human- and machine-generated text. This perspective is particularly relevant for Italian, where syntactic variation, flexible word order, and rich morphosyntactic agreement provide potential cues beyond surface fluency. Recent work suggests that LLM-generated text, especially when produced by English-centric models, may retain subtle structural biases that manifest at the syntactic level. Accordingly, we investigate whether detection signals are better captured through representations that emphasize structure rather than purely semantic abstraction.

A central modeling question therefore concerns the granularity at which such discriminative signals emerge. Some artifacts of machine-generated text may be highly localized (e.g., repetitive phrasing or short-range regularities), while others may only become apparent when considering global discourse organization and coherence. To study this trade-off, we explore two complementary training regimes. In a *local-context* setting, documents are split into overlapping chunks of three sentences, which are classified independently. This increases the number of training instances and emphasizes short-range artifacts. At inference time, chunk-level predictions are aggregated via majority voting to obtain a document-level label. In a *global-context* setting, the encoder processes the entire document, allowing the model to exploit long-range discourse structure. For full-document training, inputs are truncated to 1,024 tokens, which covers the majority of documents in our data.

Within the chunk-based regime, we further examine where discriminative information is encoded inside the model. BERT-like encoders distribute linguistic information across layers: intermediate representations often make syntactic structure more linearly accessible than the top layer [28]. We therefore evaluate features extracted not only from the final [CLS] embedding, but also from intermediate layers. This enables a controlled assessment of whether syntactic representations contribute to MGT detection. These design choices directly motivate the layer-wise ablation analysis reported in Section 4 and are grounded in the probing results presented in Appendix A.

All supervised systems follow a standard encoder-classifier architecture. A pretrained transformer encoder processes the input text, and a lightweight classification head predicts a single document-level label. Our primary backbone is ModernBERT [15], an English-centric encoder optimized for long-context processing through an alternation of local and global attention layers. We additionally consider mModernBERT [29] as comparative baseline.

Finally, we investigate the effect of model capacity and pretraining. We compare ModernBERT-base and ModernBERT-large to assess whether increased capacity improves cross-lingual transfer, and contrast English-centric and multilingual encoders under the same training protocol. All supervised models are fine-tuned using cross-entropy loss.

For the official submission, we selected ModernBERT-large trained on an augmented dataset with full-document context, prioritizing robustness to unseen generators. A post-hoc analysis, discussed in Section 4.1, shows that the same architecture trained exclusively on the original dataset achieves the highest accuracy overall.

### 3.2. Subtask B: Human - Machine Text Segmentation

Subtask B addresses a more fine-grained and realistic scenario in which human-written and machine-generated text are interwoven within the same document. Given a text that begins with a human-authored prompt and continues with a machine-generated segment, the goal is to identify the character index at which the transition occurs. Unlike Subtask A, which focuses on document-level classification, this setting requires precise boundary localization. As in Subtask A, we investigate both a data-agnostic approach and a supervised method to analyze their effectiveness under these conditions.

#### 3.2.1. Unsupervised Setting: Token-Level Fast-DetectGPT Adaptation

As a data-agnostic baseline for Subtask B, we reuse Fast-DetectGPT [13], introduced in Section 3.1.1, and investigate whether its detection signal can be repurposed for boundary localization. All experiments employ the same scoring model selected for Subtask A, namely Llama-3.2-8B-Instruct, to ensure consistency across tasks. In Subtask A, Fast-DetectGPT aggregates token-level discrepancy scores into a single document-level statistic to determine whether a text is human- or machine-generated. In Subtask B, however, each document is composed of both human-authored and machine-generated segments. This setting makes a global aggregation unsuitable, as the detection signal is expected to vary within the same document. A natural extension is therefore to inspect the Fast-DetectGPT discrepancy signal locally at the token level.

Our intuition is that mixed human-machine text should exhibit two distinct statistical regimes: an initial human-authored region characterized by near-zero discrepancy scores, followed by a machine-generated region with consistently positive values. While we cannot use those individual values to predict the exact splitting point, we hypothesize that we can use those aggregated scores to find an area in which the slope of the cumulative discrepancy score sharply increases, and use that as a boundary detection estimator.

To make this intuition operational, we retain the original Fast-DetectGPT discrepancy score  $s_j$  formulation. Instead of aggregating only at the end of the paragraph, for each token position  $j$  we compute the relative cumulative discrepancy score:

$$s_j = \log P(x_j \mid x_{<j}) + H_j \quad (1)$$

$$H_j = - \sum_{v \in \mathcal{V}} P(v \mid x_{<j}) \log P(v \mid x_{<j}) \quad (2)$$

Where  $\log P(x_j \mid x_{<j})$  is the log-probability of the observed token given the preceding context, and  $H_j$  is the entropy of the predictive distribution at position  $j$ . Intuitively,  $s_j$  measures how much an observed token is more likely than a randomly sampled alternative: machine-generated tokens tend to occupy high-probability regions, yielding positive scores, while human-authored tokens, even cumulated, center around  $\approx 0$  [14].

Since boundary localization requires stability over noise, we aggregate these scores cumulatively using a normalized cumulative statistic in which where  $\sigma_i^2$  is the variance of  $\log P(v \mid x_{<i})$  under the model’s predictive distribution:

$$Z_j = \frac{\sum_{i=0}^j s_i}{\sqrt{\sum_{i=0}^j \sigma_i^2}} \quad (3)$$

We rely on heuristic strategies to infer the transition point from the normalized cumulative trajectory  $Z_j$ . Since we are looking at the point in which the slope changes, we propose a combined score method that leverages both the first derivative (slope) and second derivative (curvature) of the smoothed cumulative trajectory  $Z_j$ . The trajectory is first smoothed using a Gaussian filter to reduce noise, after which both derivatives are computed and independently normalized to the  $[0, 1]$  range to ensure comparability. The final decision signal is an average of the normalised first and second derivatives.

The predicted boundary is the position maximizing this combined signal, excluding a margin of 10% at both ends to avoid spurious detections near document boundaries, respecting the peculiarity of the challenge data set, in which the tails are always human.

### 3.2.2. Supervised Setting: Token-Level Classification

For the supervised setting of Subtask B, we address boundary detection by directly learning token-level cues from data. While the unsupervised approach relies on word-level predictability estimated by a language model, here we ask whether similar information can be learned implicitly by an encoder trained on labeled examples.

We fine-tune a transformer encoder to perform binary token-level classification, where each token is labeled as either human-authored (0) or machine-generated (1). The intuition mirrors the unsupervised case: machine-generated continuations tend to be locally more regular and predictable, while human text exhibits higher variability. However, predicting whether a single token corresponds to the exact transition point is an extremely challenging objective. As discussed in the unsupervised setting, individual tokens are often highly constrained by syntax and discourse, regardless of authorship.

Crucially, the output of the encoder is not a hard decision but a probability distribution. For each token  $t_i$ , the model produces a probability  $p_i \in [0, 1]$  indicating the likelihood that the token belongs to the machine-generated segment. As in BERT-style encoders, these probabilities are context-dependent and reflect information aggregated over the surrounding tokens. While the classifier is not expected to be reliable exactly at the boundary, it can reliably identify regions that are clearly human-written or clearly machine-generated.

For this reason, boundary detection is not performed by thresholding individual token predictions. Moreover, the heuristics used in the unsupervised setting are not applicable here, as they are defined over word-level probability distributions rather than learned classification scores. Instead, we frame boundary localization as a change point detection problem over the sequence of predicted probabilities.

Given a sequence of  $N$  tokens with associated probabilities  $\{p_1, \dots, p_N\}$ , we score each candidate boundary position  $k$  using the following change point detection formula:

$$\text{Score}(k) = \sum_{i=1}^k \log(1 - p_i) + \sum_{j=k+1}^N \log(p_j) \quad (4)$$

The first term accumulates evidence that tokens before  $k$  are human-authored, while the second term accumulates evidence that tokens after  $k$  are machine-generated. The predicted boundary is selected as:

$$k^* = \arg \max_k \text{Score}(k)$$

This formulation favors split points where confident human predictions are concentrated on the left and confident machine predictions on the right, while uncertain tokens near the boundary have limited influence. We employ ModernBERT as the primary encoder, augmented with a two-layer MLP classification head. As a follow-up analysis, we replicate the same training procedure using `mModernBERT` to assess the impact of multilingual pretraining under limited Italian supervision. Token indices are finally mapped to character offsets by selecting the first character of the predicted boundary token.

## 4. Results and Discussion

This section details the experimental evaluation of the systems described in the Section 3. We structure the analysis by subtask, first defining the experimental protocol and baselines, and then discussing the results with a distinction between the official submission and post-hoc analyses. Finally, we synthesize our findings regarding the broader research questions on robustness and representation depth.



## 4.1. Subtask A: Machine-Generated Text Detection

In this section, we report results on the official blind test set of the shared task, in which both domains and generators are unseen during training. We compare supervised encoders with zero-shot baselines, analyze the effects of data augmentation and context size, and discuss which architectural choices support effective cross-lingual detection.

### 4.1.1. Experimental Setup

**Data Distribution.** Experiments were conducted on the official DeSegMa-IT training dataset. While the class distribution is perfectly balanced (50% human, 50% machine), we observe a significant *length asymmetry*: human-authored texts are on average twice as long as machine-generated ones (799 vs 406 tokens) and exhibit much higher variance. This “brevity bias” acts as a potential confounder, as a model might learn to associate shortness with artificiality. To improve robustness against unseen generators, we adopted an augmented training regime by injecting 14,000 additional instances: 7,000 human articles from the CHANGE-IT<sup>2</sup> corpus of Italian newspaper articles and 7,000 synthetic samples generated by translating the English RAID benchmark [19]<sup>3</sup> via Qwen2.5-7B-Instruct.

The blind test-set of the is composed by 15,135 instances.

**Investigated Systems.** To disentangle the contribution of architecture, context size, and training data, we benchmark a comprehensive set of configurations corresponding to the entries in Table 1:

- **Unsupervised Baseline (Fast-DetectGPT):** We employ Fast-DetectGPT with Llama-3.2-8B-Instruct as the scoring model. This serves as a data-agnostic reference point, quantifying how much detection signal is available purely from distributional artifacts (log-likelihood curvature) without access to labeled training examples.
- **Supervised Backbone (ModernBERT):** Our primary analysis relies on ModernBERT-large, evaluated under three distinct regimes to test robustness:
  1. *Local:* The model classifies independent 3-sentence windows, aggregating predictions via majority vote<sup>4</sup>. This tests detection reliability in low-context scenarios.
  2. *Global:* The model processes the entire document (up to 8,192 tokens) to capture long-range discourse incoherence. This represents our strongest configuration trained on the official data.
  3. *Global + Aug (Submitted):* The official submission configuration, identical to the Global setting but trained on the augmented dataset (including RAID Qwen-translated texts) to test generalization to unseen generators.
- **Comparative Architectures:** To validate whether ModernBERT’s performance stems from its architecture or cross-lingual transfer, we compare it against **mModernBERT**: the multilingual variant of our primary backbone. Comparing this against the English-centric ModernBERT allows us to isolate the impact of language-specific pretraining versus structural transfer.
- **Most Frequent Class (MFC):** A trivial baseline ( $\approx 50.10\%$ ) included solely to establish the random-guess floor for the balanced test set.

The selection of the submitted configuration (ModernBERT Large with Data Augmentation) was necessitated by the lack of a discriminative signal on our internal validation set. During the development phase, we observed a pronounced ceiling effect, where several explored configurations achieved near-perfect performance, rendering them empirically indistinguishable. Consequently, we prioritized the system trained with Data Augmentation based on the hypothesis that exposure to a more diverse data

<sup>2</sup><https://huggingface.co/datasets/MattiaSangermano/change-it>

<sup>3</sup><https://raid-bench.xyz/>

<sup>4</sup>Technically, we employ a *soft-voting* aggregation strategy. Rather than counting discrete class labels, we extract the softmax probability  $P(\text{Machine} \mid c_i)$  vs.  $P(\text{Human} \mid c_i)$  and use them for weighted voting.

distribution would bring superior generalization capabilities and robustness against distribution shifts in the blind test set.

**Metrics.** Primary evaluation is based on Accuracy, the official metric for Subtask A. Additionally, we report Precision, Recall, and Macro-F1 to provide a granular view of error types and class-specific behavior.

#### 4.1.2. Main Results

Table 1 presents the results on the official test set. The table shows our unsupervised model based on Fast-DetectGPT and the naive baseline in the first two rows. Inside rows 3-5 we display our supervised models based on the ModernBERT architecture. On row 6 we present the best competitor system for reference.

The hierarchy of performance provides clear answers to our research questions regarding the trade-off between robustness and adaptation.

**Supervised vs. Unsupervised.** The first major finding is the substantial gap between data-agnostic and supervised methods. Fast-DetectGPT achieves a respectable accuracy of 82.86%, confirming that machine-generated texts in Italian leave detectable statistical traces (e.g., negative curvature regions in log-likelihood) that persist even zero-shot. However, supervised adaptation yields a relative error reduction of over 75% (best supervised model: 95.78%). This suggests that while generic statistical artifacts are present, the specific generators used in the wild exhibit idiosyncratic patterns that are best captured through fine-tuning.

**The Failure of Naive Augmentation.** A counter-intuitive result concerns our submission strategy. Our official run, trained on the *Augmented* dataset (which included translated RAID data and Qwen-generated texts), reached 92.87% accuracy. However, our post-hoc analysis reveals that training *exclusively* on the official provided data (Original) would have yielded significantly higher performance (95.78%), outperforming the best competitor system (94.57%). This indicates that the augmentation strategy introduced a distributional shift that was detrimental rather than helpful. The specific stylistic signatures of the "in-the-wild" test set were likely well-represented in the original training data. By injecting synthetic data from different domains and generators (e.g., Qwen), we likely diluted the model's focus, validating the hypothesis that MGT detection is highly sensitive to the specific generator-domain pairs used for training.

**Context Matters: Local Context vs. Global Context.** We compare processing granularity by contrasting a chunk-based approach (voting on 3-sentence windows) against a full-document approach. The full-document model outperforms the chunked variant by nearly 10 percentage points (95.78% vs 85.97%). This result has a crucial theoretical implication: detection signals are not merely local glitches (e.g., repetition or non-sequiturs visible in a short window) but are deeply rooted in global discourse coherence. The model requires long-range attention to detect the subtle "fading" of logic or structure typical of neural generation.

**Architectural Analysis: Cross-Lingual Transfer.** A striking observation from Table 1 is the performance of the English-centric ModernBERT-large (95.78%) compared to its multilingual counterpart mModernBERT (92.82%). Conventional wisdom suggests that for Italian tasks, a multilingual model should be superior. Why does an English-centric model win? We argue that, at document-level, MGT detection relies less on language-specific lexical knowledge (vocabulary coverage) and more on *structural* and *distributional* features (e.g., entropy plateaus, syntactic regularity) which are artifacts of the Transformer architecture itself. Since LLMs (the generators) share similar underlying architectures regardless of the output language, the artifacts they introduce are likely universal. ModernBERT's

**Table 1**

Subtask A results on the **official test set**. All values are reported in percentages (%). **Submitted** denotes our official competition entry. The *Best Competitor System* corresponds to the first-place submission by team *Gradient Descenders*. *MFC* denotes the Most Frequent Class baseline.

Run	Model	Acc	P	R	F1
Unsupervised	MFC (Most Frequent Class Baseline)	50.10	25.05	50.00	33.38
	Fast-DetectGPT (scorer: LLama3 . 2 - 8B)	82.86	84.68	82.88	82.64
Supervised	ModernBERT-large (local)	85.97	89.02	85.97	87.47
	mModernBERT (global)	92.82	93.66	92.78	92.78
	ModernBERT-large (global + aug) ( <b>Submitted</b> )	92.87	93.63	92.85	92.82
	ModernBERT-large (global)	<b>95.78</b>	95.99	95.78	95.77
NA	Best Competitor System	94.57	-	-	-

superior pre-training on code and massive English corpora likely allowed it to learn more robust representations of "artificiality" and structural coherence, which transferred zero-shot to Italian syntax more effectively than the noisy multilingual alignment of mBERT variants.

**The "Safe Zone" of Detection.** Table 2 breaks down performance by document length. The results highlight a clear boundary condition for encoder-based detectors. On documents longer than 512 tokens, the model is effectively infallible (99.9% accuracy). This confirms that given sufficient statistical evidence, the distinction between human and machine text is sharp.

**The Short-Text Failure Mode.** Conversely, performance drops drastically to 76.8% for texts shorter than 128 tokens. This correlates with the "Brevity Bias" observed in the training data: the model struggles to disentangle "shortness" from "artificiality," or simply lacks the necessary context window to observe discourse-level incoherence. This remains the primary open challenge for future research: detecting MGT in low-resource (short context) regimes requires signal amplification strategies that standard fine-tuning does not provide.

**Table 2**

Subtask A: Error Analysis by Document Length (ModernBERT-large). The number of tokens is calculated based on the ModernBERT tokenizer.

Text Length (tokens)	Accuracy(%)
Short (< 128)	76.83
Medium (128–512)	89.95
Long (> 512)	<b>99.90</b>

**Layer-Wise Ablation Study.** To determine whether MGT detection relies primarily on low-level structural artifacts or requires high-level semantic abstraction, we conducted a layer-wise ablation study using ModernBERT-base in the local-context (chunked) setting. Results are detailed in Table 4 in appendix A. Contrary to the hypothesis that detection signals saturate in intermediate syntactic layers, we observe a monotonic improvement in performance throughout the network depth. While the first layer already provides a strong baseline (Accuracy: 81.23%), indicating the presence of surface-level lexical or morphological cues, reliance on intermediate representations (Layers 11–14) yields only partial gains (84.63%). The highest performance is achieved strictly at the final classification layer (87.15%). This trajectory suggests that while "glitches" in machine-generated text may begin at the surface level, reliable detection of advanced generators requires the full compositional capacity of the Transformer to model complex discourse-level dependencies.

## 4.2. Subtask B: Human–Machine Text Segmentation

### 4.2.1. Experimental Setup

**Data Distribution.** Experiments focus on the training and test split of the DeSegMa-IT Subtask B dataset, comprising 19,945 mixed-source documents for the training section and of 23,211 documents for the test section.

Each instance consists of a human-authored prompt followed by a machine-generated continuation. An analysis of the training distribution reveals a critical structural property: the transition point is located, on average, at character 264 ( $\pm 116.9$ ). This implies that the human context is relatively short (averaging 84.5 tokens), while the machine-generated continuation is dominant (averaging 355.0 tokens). This asymmetry poses a significant modeling challenge: the encoder must detect the stylistic shift relying on a very limited "warm-up" window of human text to establish a baseline. Finally, the total document length in the training set remains well within the capacity of ModernBERT’s context window, ensuring that boundary artifacts are not conflated with truncation noise during fine-tuning.

**Investigated Systems.** To isolate the contributions of supervision and multilingual pretraining to boundary localization, we benchmark the following configurations corresponding to the entries in Table 3:

- **Unsupervised Baseline (Token-Level Fast-DetectGPT):** We adapt Fast-DetectGPT to the segmentation setting by retaining its token-level discrepancy scores rather than aggregating them into a single document statistic. Using Llama-3.1-8B-Instruct as the scoring model (consistent with Subtask A), we compute normalized cumulative trajectories and apply a combined first- and second-derivative heuristic to detect the point of maximum slope change, as described in Section 3.1.1. This serves as a data-agnostic reference point, testing whether local probability curvature alone can signal authorship transitions.
- **Supervised Backbone (ModernBERT-large, Submitted):** Our official submission employs 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, and final boundary prediction is obtained via the change point detection procedure described in Equation (4). Token indices are mapped to character offsets by selecting the first character of the predicted boundary token.
- **Multilingual Variant (mModernBERT):** To assess whether fine-grained segmentation benefits from native morphological representations, we replicate the supervised training procedure using mModernBERT-base. This comparison isolates the impact of multilingual pretraining versus cross-lingual transfer from an English-centric encoder.
- **Middle Split Baseline:** A naive baseline that places the boundary at 50% of the document length, included to establish the difficulty floor for the localization task.

**Metrics** Evaluation is performed using Mean Absolute Error (MAE), which measures the average character distance between the predicted boundary and the ground truth. Unlike binary classification metrics, MAE quantifies the precision of localization, penalizing errors proportionally to their magnitude. To benchmark the difficulty of the task, we introduce a naive Middle-Split Baseline, which blindly places the boundary at 50% of the document length.

### 4.2.2. Main Results

Table 3 summarizes Subtask B performance measured by Mean Absolute Error (MAE) in characters. In the first two rows is presented a naive splitting at the 50% mark to use as reference and our unsupervised Token-level Fast-DetectGPT. On rows 3-4 we display our supervised models based on the ModernBERT architecture employing the change point detection formula 3.2.2. On row 5 we present the best system presented in the challenge as a point of reference.

Our comparative analysis targets two specific dimensions of the segmentation problem. First, we assess the viability of Zero-Shot Curvature Estimation by applying Token-Level Fast-DetectGPT, testing whether local perplexity spikes are sufficient to signal authorship changes without supervision. The token-level Fast-DetectGPT adaptation achieves 158.31 MAE, a substantial improvement over the middle-split baseline (695.37) but nearly double the error of our submitted supervised system. This gap suggests that local probability curvature, while informative at the document level, provides insufficient granularity for precise boundary localization without a different heuristic compared to the usage of derivatives or additional syntactic informations.

Second, within the supervised framework, we investigate the role of Multilingual Alignment. By comparing the English-centric ModernBERT (our official submission) against the multilingual mModernBERT (post-hoc), testing the hypothesis that fine-grained segmentation, unlike document-level detection, requires the native morphological representations offered by multilingual pretraining.

Our submitted ModernBERT-large system achieves 81.6 MAE, corresponding to an average localization error of approximately 13 words, in line with our a-priori evaluation on an held-out portion of the training set. The post-hoc experiments with mModernBERT yield a 25.7% relative improvement, reducing MAE to 60.58 characters and narrowing the gap with the best competitor to approximately 8 characters. This result contrasts with Subtask A, where ModernBERT-large outperformed MultilingualModernBERT-base by almost 3%. We attribute this reversal to the differing task demands: token-level boundary detection requires fine-grained morphosyntactic sensitivity that multilingual pre-training better provides, whereas document-level classification benefits from global coherence signals that transfer well cross-lingually.

**Table 3**

Subtask B results on the **official test** (MAE in characters; lower is better). **Submitted** indicates the official run.

Run	Model	MAE
Unsupervised	Middle Split Baseline	695.37
	Token-level Fast-DetectGPT (scorer: LLama3 . 2 - 8B)	158.31
Supervised	ModernBERT + Change-Point ( <b>Submitted</b> )	81.60
	mModernBERT + Change-Point	60.58
NA	Best Competitor System	52.54

## 5. Conclusions

This work compared supervised and unsupervised approaches for Italian machine-generated text detection on the DeSegMa-IT benchmark. For Subtask A (document-level classification), our official submission, ModernBERT-large with data augmentation and full-document context, achieved 92.87% accuracy, placing third in the challenge. The same ModernBERT-large trained without data-augmentation for three epochs achieved 95.78% accuracy in a successive ablation study, while MultilingualModernBERT-base scored 92.82%. Layer-wise ablations on ModernBERT-base using chunk-based classification reveal that intermediate layers (11–14) achieve 84.6% accuracy compared to 87.2% for the final [CLS] embedding, suggesting that syntactic representations in the encoder’s middle layers already capture most of the discriminative signal. For Subtask B (boundary detection), the supervised ModernBERT-large system with change point detection obtained 81.6 MAE, which improved to 60.6 MAE when replaced with the multilingually pre-trained MultilingualModernBERT-base. This reversal shows that while *model size* is the decisive factor for Subtask A, *multilingual pretraining* is more relevant in the context of Subtask B. This confirms our hypothesis that MGT detection at the document level relies more on structural features that generalize across languages, while fine-grained segmentation requires native semantic and morphological knowledge.

Our submitted model scored almost perfectly - 99.9%+ accuracy - on texts longer than 512 tokens, dropping to 76.83% for short texts under 128 tokens, with almost all errors being false negatives. If



those findings were replicated consistently on different text-domains and against adversarial prompting and ablated models, automatic MGT detectors can be suggested for usage on long texts.

For both experiments the unsupervised domain-agnostic systems, predictably, scored behind their supervised counterparts. On Subtask A, Fast-DetectGPT using LLama3.2-8B as evaluator reached 84%, while the token-level Fast-DetectGPT adaptation yielded 158 MAE on Subtask B, suggesting that local probability curvature alone is insufficient for precise boundary localization. Contrary to our expectation, adding 14,000 new documents coming from RAID [19] as Data Augmentation for Subtask A reduced by almost 3% the model performance. This suggests that augmentation with translated data introduces distributional artifacts that outweigh robustness gains on this benchmark. Since FastDetectGPT makes no prior assumptions about the language or domain, we expect its performance to remain flat across domains, registers, and languages, while supervised model performance should decrease with respect to distance to the training dataset. This makes Fast-DetectGPT, and unsupervised methods in general, paradoxically more suited to real-world application than benchmarks. Prior work has shown that multilingual LLMs retain English-influenced syntactic patterns even when generating in other languages [20], and our probing analysis (Appendix A) reveals that hierarchical depth information remains weakly encoded in ModernBERT. We believe that incorporating explicit syntactic features alongside semantic signals offers currently the most promising avenue for exploration.

## Acknowledgments

The authors acknowledge financial support from the PNRR MUR project PE0000013-FAIR. Moreover, the authors acknowledge support from Project ECS 0000024 Rome Technopole, - CUP B83C22002820006, NRP Mission 4 Component 2 Investment 1.5, Funded by the European Union - NextGenerationEU.

## Declaration on Generative AI

Parts of the writing and editing process for this manuscript were supported by the use of generative AI tools, specifically Google’s Gemini. These tools were employed to enhance clarity and correct spelling. All AI-assisted content was carefully reviewed and validated by the authors to ensure accuracy, originality, and compliance with ethical and scientific standards. The authors bear full responsibility for the final content.

## References

- [1] A. Liu, A. Mei, B. Lin, B. Xue, B. Wang, B. Xu, B. Wu, B. Zhang, C. Lin, C. Dong, et al., Deepseek-v3. 2: Pushing the frontier of open large language models, arXiv preprint arXiv:2512.02556 (2025).
- [2] K. Team, Y. Bai, Y. Bao, G. Chen, J. Chen, N. Chen, R. Chen, Y. Chen, Y. Chen, Y. Chen, et al., Kimi k2: Open agentic intelligence, arXiv preprint arXiv:2507.20534 (2025).
- [3] A. Ardit, O. Obeso, A. Syed, D. Paleka, N. Panickssery, W. Gurnee, N. Nanda, Refusal in language models is mediated by a single direction (2024). arXiv:2406.11717.
- [4] A. K. Kadhim, L. Jiao, R. A. Shafik, O.-C. Granmo, Adversarial attacks on ai-generated text detection models: A token probability-based approach using embeddings, 2025. URL: <https://arxiv.org/abs/2501.18998>. arXiv:2501.18998, preprint.
- [5] Y. Zhou, B. He, L. Sun, Humanizing machine-generated content: Evading AI-text detection through adversarial attack, in: N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, N. Xue (Eds.), Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), ELRA and ICCL, Torino, Italia, 2024, pp. 8427–8437. URL: <https://aclanthology.org/2024.lrec-main.739/>.
- [6] A. Uchendu, Z. Ma, T. Le, R. Zhang, D. Lee, TURINGBENCH: A benchmark environment for Turing test in the age of neural text generation, in: M.-F. Moens, X. Huang, L. Specia, S. W.-t. Yih (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2021, Association for

- Computational Linguistics, Punta Cana, Dominican Republic, 2021, pp. 2001–2016. URL: <https://aclanthology.org/2021.findings-emnlp.172/>. doi:10.18653/v1/2021.findings-emnlp.172.
- [7] B. Guo, X. Zhang, Z. Wang, M. Jiang, J. Nie, Y. Ding, J. Yue, Y. Wu, How close is chatgpt to human experts? comparison corpus, evaluation, and detection, 2023. URL: <https://arxiv.org/abs/2301.07597>. arXiv:2301.07597.
  - [8] Y. Wang, J. Mansurov, P. Ivanov, J. Su, A. Shelmanov, A. Tsvigun, O. Mohammed Afzal, T. Mahmoud, G. Puccetti, T. Arnold, SemEval-2024 task 8: Multidomain, multimodel and multi-lingual machine-generated text detection, in: A. K. Ojha, A. S. Doğruöz, H. Tayyar Madabushi, G. Da San Martino, S. Rosenthal, A. Rosá (Eds.), Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), Association for Computational Linguistics, Mexico City, Mexico, 2024, pp. 2057–2079. URL: <https://aclanthology.org/2024.semeval-1.279/>. doi:10.18653/v1/2024.semeval-1.279.
  - [9] Y. Wang, J. Mansurov, P. Ivanov, J. Su, A. Shelmanov, A. Tsvigun, C. Whitehouse, O. M. Afzal, T. Mahmoud, T. Sasaki, et al., M4: Multi-generator, multi-domain, and multi-lingual black-box machine-generated text detection, in: Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), 2024, pp. 1369–1407.
  - [10] G. Sarti, M. Nissim, It5: Text-to-text pretraining for italian language understanding and generation, in: International Conference on Language Resources and Evaluation, 2022. URL: <https://api.semanticscholar.org/CorpusID:247315276>.
  - [11] G. Puccetti, A. Pedrotti, A. Esuli, Desegma-it at evalita 2026: Overview of the detection and segmentation of machine generated text in italian task, 2026.
  - [12] 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.
  - [13] G. Bao, Y. Zhao, Z. Teng, L. Yang, Y. Zhang, Fast-detectgpt: Efficient zero-shot detection of machine-generated text via conditional probability curvature, in: Proc. of ICLR, 2023.
  - [14] 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.
  - [15] B. Warner, A. Chaffin, et al., Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference, in: W. Che, J. Nabende, E. Shutova, M. T. Pilehvar (Eds.), Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Vienna, Austria, 2025, pp. 2526–2547. URL: <https://aclanthology.org/2025.acl-long.127/>. doi:10.18653/v1/2025.acl-long.127.
  - [16] S. T. Lekkala, Y. Annepaka, A. K. Challa, S. R. Machireddy, P. Pakray, C. Chunka, Mixture of detectors: A compact view of machine-generated text detection, arXiv preprint arXiv:2509.22147 (2025).
  - [17] J. Pu, Z. Sarwar, S. M. Abdullah, A. Rehman, Y. Kim, P. Bhattacharya, M. Javed, B. Viswanath, Deepfake text detection: Limitations and opportunities, in: 2023 IEEE Symposium on Security and Privacy (SP), 2023, pp. 1613–1630. doi:10.1109/SP46215.2023.10179387.
  - [18] F. Borazio, C. D. Hromei, E. Passone, D. Croce, R. Basili, Mm-iglu-it: Multi-modal interactive grounded language understanding in italian, in: A. Artale, G. Cortellessa, M. Montali (Eds.), AIXIA 2024 – Advances in Artificial Intelligence, Springer Nature Switzerland, Cham, 2025, pp. 64–78.
  - [19] L. Dugan, A. Hwang, F. Trhlík, A. Zhu, J. M. Ludan, H. Xu, D. Ippolito, C. Callison-Burch, RAID: A shared benchmark for robust evaluation of machine-generated text detectors, in: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 12463–12492. URL: <https://aclanthology.org/2024.acl-long.674>.
  - [20] Y. Guo, S. Conia, Z. Zhou, M. Li, S. Potdar, H. Xiao, Do large language models have an english accent? evaluating and improving the naturalness of multilingual llms, in: Proceedings of the

- 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2025, pp. 3823–3838.
- [21] E. S. Ruzzetti, F. Rinaldi, D. Onorati, D. Venditti, L. Rinaldi, T. Caselli, F. M. Zanzotto, Assessing the asymmetric behaviour of italian large language models across different syntactic structures, in: Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024), 2024, pp. 854–863.
  - [22] D. Macko, J. Kopal, R. Moro, I. Srba, Multisocial: Multilingual benchmark of machine-generated text detection of social-media texts, in: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2025, pp. 727–752.
  - [23] O. Al Minshidawi, A.-H. Vahabie, Classifying ai-generated text in low-resource languages like arabic, *AUT Journal of Modeling and Simulation* 57 (2025) 113–124.
  - [24] Y. Wang, J. Mansurov, P. Ivanov, J. Su, A. Shelmanov, A. Tsvigun, O. Mohammed Afzal, T. Mahmoud, G. Puccetti, T. Arnold, A. Aji, N. Habash, I. Gurevych, P. Nakov, M4GT-bench: Evaluation benchmark for black-box machine-generated text detection, 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. 3964–3992. URL: <https://aclanthology.org/2024.acl-long.218/>. doi:10.18653/v1/2024.acl-long.218.
  - [25] B. Magnini, R. Zanoli, M. Resta, M. Cimmino, P. Albano, M. Madeddu, V. Patti, Evalita-LLM: Benchmarking large language models on italian, *arXiv preprint arXiv:2502.02289* (2025). [arXiv:2502.02289](https://arxiv.org/abs/2502.02289).
  - [26] L. Moroni, A. Esuli, D. Marcheggiani, B. Magnini, Towards a more comprehensive evaluation for italian LLMs, in: Proceedings of the 2024 Workshop on Computational Linguistics for Italian (CLiCit 2024), 2024, pp. 48–59.
  - [27] A. Esuli, F. Falchi, M. Malvaldi, G. Puccetti, You write like a GPT, in: F. Dell’Orletta, A. Lenci, S. Montemagni, R. Sprugnoli (Eds.), Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024), CEUR Workshop Proceedings, Pisa, Italy, 2024, pp. 343–348. URL: <https://aclanthology.org/2024.clicit-1.41/>.
  - [28] J. Hewitt, C. D. Manning, A structural probe for finding syntax in word representations, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4129–4138. URL: <https://aclanthology.org/N19-1419/>. doi:10.18653/v1/N19-1419.
  - [29] M. Marone, O. Weller, W. Fleshman, E. Yang, D. Lawrie, B. Van Durme, mmbert: A modern multilingual encoder with annealed language learning, *arXiv preprint arXiv:2509.06888* (2025). URL: <https://arxiv.org/abs/2509.06888>.
  - [30] L. Schut, Y. Gal, S. Farquhar, Do multilingual llms think in english?, in: ICLR 2025 Workshop on Building Trust in Language Models and Applications, 2025.
  - [31] A. Conneau, G. Kruszewski, G. Lample, L. Barrault, M. Baroni, What you can cram into a single \$&!#\* vector: Probing sentence embeddings for linguistic properties, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 2126–2136. URL: <https://aclanthology.org/P18-1198>. doi:10.18653/v1/P18-1198.
  - [32] J. Hewitt, P. Liang, Designing and interpreting probes with control tasks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 2733–2743. URL: <https://aclanthology.org/D19-1275>. doi:10.18653/v1/D19-1275.
  - [33] G. Jawahar, B. Sagot, D. Seddah, What does BERT learn about the structure of language?, in: A. Korhonen, D. Traum, L. Màrquez (Eds.), Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 3651–3657. URL: <https://aclanthology.org/P19-1356/>. doi:10.18653/v1/P19-1356.

## A. Layer Ablation and Probing Analysis

We provide here the grounding and detailed analysis for the layer-wise ablation study briefly discussed in Section 4.1.2. To determine which internal representations are most discriminative for Italian MGT detection, we evaluated ModernBERT-base using three distinct configurations (Table 4): the first layer, a specific block of intermediate layers, and the final output embedding. This investigation reveals that while detection performance improves with depth, distinct linguistic signals contribute differently to the model’s decision boundary.

**Table 4**

Ablation study on representation depth and model size using local context (3-sentence chunks). All values are reported in percentages (%). Performance consistently improves with layer depth, confirming that high-level semantic features (Final Layer) are more discriminative than surface or syntactic ones.

Model	Configuration	Acc	P	R	F1
ModernBERT-base	Layer 1 (Mean)	81.23	86.29	81.23	83.68
ModernBERT-base	Intermediate Layers 11-14	84.63	88.18	84.63	86.37
ModernBERT-base	Final Layer [CLS]	<b>87.15</b>	<b>89.70</b>	<b>87.15</b>	<b>88.41</b>

We selected **Layer 1** to test the hypothesis that MGT detection might be solvable via low-level surface artifacts alone. Using this layer yields an accuracy of 81.23%. While significantly above random chance, it lags behind deeper representations, suggesting that while surface signals are computationally cheap to extract, they are insufficient for high-accuracy detection.

The selection of **Layers 11-14** (84.63%) was derived from a specific linguistic hypothesis regarding “English-mediated” generation. Recent work on multilingual LLMs suggests that outputs are often syntactically “English-accented” (reflecting English) influenced grammatical preferences and distributional artifacts even when the surface language is Italian. Guo et al. [20] demonstrate that this accent manifests as systematic shifts in dependency relations and hierarchical substructures (e.g., tree configurations), while mechanistic studies suggest an English-centered latent space for semantic decisions [30]. If machine-generated text exhibits such structural rigidity, a detector should be able to recover these signals from the encoder’s intermediate representations.

To investigate the internal linguistic representations of ModernBERT-base, we adopt the layer-wise probing methodology of Conneau et al. [31] and evaluate seven diagnostic tasks spanning three linguistic dimensions<sup>5</sup>. For surface features, we use (1) *Length*, which predicts the total number of words to assess retention of simple information, and (2) *BigramShift*, which tests sensitivity to local word order by detecting whether adjacent tokens have been swapped. For syntactic features, we include tasks targeting both *vertical* and *horizontal* structure: (3) *Depth*, which estimates the maximum depth of the syntactic parse tree as a measure of hierarchical complexity, and (4) *TopConst*, which predicts the sequence of top-level constituents (e.g., NP–VP) to capture global syntactic organization. Finally, for semantic features, we consider (5) *Tense*, which identifies the grammatical tense of the main verb (e.g., past vs. present); (6) *SubjNum* and (7) *ObjNum*, which detect the grammatical number (singular/plural) of the subject and direct object, respectively.

Using the *Control Task* framework to ensure selectivity [32], our analysis revealed a clear functional stratification within ModernBERT, as shown in Table 5. While surface features like sentence length are accessible throughout the network and semantic tasks (*Tense*, *SubjNum*) peak in the final layers, it’s possible to identify a distinct “Syntactic Block” in the middle layers (11–14) where structural performance maximizes, similarly to organization of the original BERT model [33].

This middle block excels at *TopConst* (peaking at Layer 14 with selectivity 0.744), suggesting strong access to coarse phrase-structural configuration. This explains why extracting representations from Layers 11–14 outperforms Layer 1: the model successfully captures the linear organization of constituents, a key dimension where translation artifacts often appear. However, the probing analysis also highlights

<sup>5</sup><https://github.com/facebookresearch/SentEval>

**Table 5**

Full Layer-wise Probing Results (Selectivity) for ModernBERT-base across all 22 layers. Values represent *Selectivity*, defined as the difference between the probe’s accuracy on the target linguistic task and a randomized control task, that is  $S = Acc_{real} - Acc_{control}$ . Note the distinct functional stratification: **Surface features** (i.e. the Text Length) impact positively ; **Syntactic features** (TopConst) show a peak in the middle block (Layers 11-14), justifying our ablation choice; while **Semantic features** (Tense, SubjNum) generally maximize in the deeper layers (18+).

Layer	Surface		Syntax		Semantics		
	<i>Length</i>	<i>BigramShift</i>	<i>Depth</i>	<i>TopConst</i>	<i>Tense</i>	<i>SubjNum</i>	<i>ObjNum</i>
0	0.767	-0.004	0.198	0.680	0.370	0.329	0.333
1	0.759	0.014	0.207	0.680	0.370	0.334	0.329
2	0.740	0.231	0.225	0.714	0.365	0.334	0.322
3	0.735	0.275	0.233	0.723	0.368	0.346	0.329
4	0.721	0.262	0.232	0.719	0.363	0.336	0.333
5	0.745	0.277	0.231	0.723	0.362	0.337	0.333
6	0.754	0.296	0.237	0.730	0.377	0.352	0.334
7	0.748	0.294	0.235	0.732	0.377	0.347	0.342
8	0.747	0.312	0.246	0.735	0.382	0.353	0.342
9	0.757	0.326	0.256	0.735	0.385	0.368	0.337
10	0.786	0.324	0.257	0.734	<b>0.390</b>	0.363	0.345
11	<b>0.793</b>	0.338	0.255	0.738	<b>0.390</b>	0.364	0.346
12	0.789	0.345	0.267	0.741	0.387	0.356	0.338
13	0.788	0.352	0.267	0.743	0.385	0.352	0.344
14	0.784	0.362	0.276	<b>0.744</b>	0.385	0.361	<b>0.349</b>
15	0.784	0.358	0.274	0.743	0.387	0.363	<b>0.351</b>
16	0.769	0.360	<b>0.278</b>	0.738	<b>0.391</b>	0.362	0.350
17	0.773	0.362	0.267	0.740	0.385	0.359	0.349
18	0.754	0.365	0.271	0.739	0.381	<b>0.381</b>	0.348
19	0.755	0.364	0.264	0.739	0.385	0.379	0.347
20	0.744	<b>0.374</b>	0.252	0.731	0.389	0.370	0.344
21	0.735	0.368	0.247	0.726	0.388	0.369	0.335
22	0.721	0.372	0.239	0.710	0.380	0.369	0.344

a critical limitation: the model struggles with syntactic *Depth* (Selectivity  $\approx 0.27$ ). Although selectivity for tree depth also peaks in this middle block, the absolute signal is weak. This suggests that while ModernBERT captures horizontal syntax, it does not preserve *hierarchical complexity* (vertical depth) in a form that remains linearly recoverable after mean pooling. Since human writing is often structurally “spikier” (greater variance in syntactic depth) than the smoother output of LLMs, the inability to fully leverage this vertical signal may explain why the syntactic block (84.63%) still underperforms compared to the final layer (87.15%), which aggregates syntactic and semantic information.