Evaluating Watermark Resilience in Italian Large Language Models Under Back-Translation Attacks

Elena Opreni[1,†], Alice Pietrasanta[1,†], Davide Altieri[1,†], Giulia Rizzi[1,†] and Elisabetta Fersini[1,*,†]

[1]University of Milano-Bicocca, Department of Informatics, Systems and Communication, Viale Sarca 336, 20126, Milan (Italy)

Abstract

The widespread deployment of Large Language Models (LLMs) highlights the need for robust provenance mechanisms to distinguish synthetic content from human authorship. While generation-time text watermarking has emerged as a prominent strategy, its empirical reliability against cross-lingual adversarial manipulation remains insufficiently explored, particularly for non-English languages. In this paper, we systematically evaluate the adversarial robustness of three representative watermarking paradigms, i.e., LogitsWatermark (KGW), SemaMark, and Semantic Invariant Robustness (SIR), across a suite of five Italian-centric LLMs subjected to untargeted back-translation attacks. We structure our empirical investigation along two primary axes. First, we analyze the impact of morphological typology (fusional, agglutinative, isolating) and resource availability across intermediate pivot languages. Our findings demonstrate that back-translation significantly degrades statistical watermark detectability, with low-resource fusional (Hindi) and agglutinative (Uzbek) pivots inducing the largest z-score distributional shifts. Second, we investigate the intersection between semantic preservation and statistical detectability. The results reveal severe vulnerability across all tested frameworks: discrete token-biasing schemes undergo near-total compromise up to 98.3%, whereas semantic embedding-based approaches mitigate degradation but still remain highly vulnerable (down to 20.8%). These insights underscore cross-lingual stability as a critical challenge for future text watermarking design.

Keywords

Text Watermarking, Italian Large Language Models, Back-translation attacks,

1. Introduction

The rapid proliferation of Large Language Models (LLMs) has transformed natural language generation, empowering models to produce highly plausible, human-like text at scale. However, the widespread of these generative technologies presents substantial societal and security risks, including automated misinformation campaigns, academic plagiarism, and the systematic erosion of digital content integrity. Consequently, establishing robust attribution frameworks able to distinguish machine-generated text from human authorship is still an open issue.

Watermarking represent one of the possible solutions to this problem, by embedding a hidden signature directly into the text sequence during the decoding phase [2]. An ideal watermarking mechanism guarantees that this signature remains imperceptible to human readers while remaining highly detectable by detection algorithms. Although such strategies demonstrate high efficacy in controlled monolingual settings, their resilience has been proven to be compromised by simple post-generation editing [3]. Among the adversarial strategies, an effective approach consists of a *back-translation attack*, wherein watermarked text is translated into a pivot target language and subsequently mapped back into the source language using neural machine translation engines. This loop acts as a semantic noisy channel, perturbing local token distributions, lexical selections, and syntactic configurations that contain the verification signal. However, to what extent such vulnerability depends

NL4AI 2026: Ninth Workshop on Natural Language for Artificial Intelligence, October 6-9th, 2026, Perugia, Italy [1]

*Corresponding author.

†These authors contributed equally.

e.opreni@campus.unimib.it (E. Opreni); a.pietrasanta@campus.unimib.it (A. Pietrasanta); d.altieri@campus.unimib.it (D. Altieri); giulia.rizzi@unimib.it (G. Rizzi); elisabetta.fersini@unimib.it (E. Fersini)

https://en.unimib.it/giulia-rizzi (G. Rizzi); https://en.unimib.it/elisabetta-fersini (E. Fersini)

0000-0002-0619-0760 (G. Rizzi); 0000-0002-8987-100X (E. Fersini)



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

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

published 2026-11-05

CEUR-WS.org/Vol-4315/paper15.pdf

on the pivot language remains uninvestigated, potentially creating asymmetrical vulnerabilities across high- and low-resource settings.

To bridge this gap, we systematically investigate back-translation adversarial attacks around two main research questions:

- ***RQ1:*** *To what extent does the morphological typology, i.e., fusional, agglutinative, isolating, of the intermediate pivot language impact watermark retention?*
- ***RQ2:*** *How do different text watermarking frameworks balance the trade-off between semantic preservation and statistical detectability under adversarial back-translation attacks?*

By exploring the intersection of watermarking resilience and multilingual attacks, this work tries to advance the development of more secure and reliable generative Italian language models. The paper is organized as follows. In Section 2, related works are outlined. In Section 3, the evaluation pipeline is introduced to support the evaluation of the vulnerability of watermarking strategies that can be adopted by Italian LLMs. In Section 4, the experimental settings are detailed, while in Section 5 the experimental results are discussed. Finally, in Section 6, conclusions are drawn, and future research directions are outlined.

2. Related Work

Textual watermarking methodologies are broadly dichotomized based on their operational stage within the text generation lifecycle, i.e., post-processing paradigms applied to existing text and generation-time techniques integrated directly into Large Language Models (LLMs). While post-processing approaches embed hidden signals into text while preserving semantic and syntactic fidelity, strategies at generation-time introduce detectable signals during the decoding process by mitigating the computational overhead. Watermarks can be further classified as zero-bit, i.e. detecting presence/absence, or multi-bit, i.e. modulating an n-bit payload.

Regarding the post-processing watermarking paradigms, two of the simplest approaches are represented by (1) format-based methods and (2) lexical and syntactic approaches. Format-based methods [4] rely on imperceptible structural modifications, such as manipulating Unicode space characters (e.g., U+2004) or injecting zero-width characters (e.g., U+200B, U+200C). Although computationally trivial and completely non-disruptive to text semantics, such methods are characterized by high fragility under basic reformatting operations typical of preprocessing tasks of LLMs. Another family of post-processing paradigms is composed of lexical and syntactic methods [5, 6, 7, 8], which exploit specific mappings to substitute tokens with contextually relevant synonyms via semantic vectors (e.g., Word2Vec) or masked language models (e.g., BERT). Syntactic approaches leverage structural transformations to map specific syntactic architectures to message bits.

Recent advancements focus on integrating watermarks directly into the decoding mechanisms of LLMs. We can broadly distinguish them into:

- Logits-Level Modification: The seminal KGW [9] framework introduces deterministic vocabulary partitioning into *green* and *red* lists based on prior token hashes, applying a scalar bias δ to green token logits. To counter the vulnerability of token-hash mechanisms to paraphrasing, SemaMark [10] aggregates the semantic embeddings of preceding context through a weighted pooling strategy, mapping them to a Normalized Embedding Ring (NE-Ring) to serve as a robust semantic seed. Similarly, the Semantics-Invariant Resilient (SIR) [11] framework extracts context embeddings using an auxiliary encoder (e.g., BERT) coupled with a neural network to produce continuous, uncrackable watermark logits. X-SIR [3] extends this paradigm to cross-lingual contexts by applying bilingual dictionaries to align semantically equivalent tokens into cross-lingual clusters.
- Token Sampling Adjustments: Moving beyond logit distortion, KTH [12] offers a *distortion-free* framework that aligns a shared pseudorandom key sequence with the model's native Cumulative

Distribution Function (CDF) via vocabulary permutations, preserving the exact statistical distribution of the LLM. At the sentence level, K-SemStamp [13] optimizes the hyperplanes of early sentence-based models by leveraging k-means clustering within the semantic embedding space, applying rejection sampling to enforce green-list cluster trajectories.

- Reinforcement-based training: The Reinforcement Watermark paradigm avoids generation-time overhead by using Proximal Policy Optimization (PPO) to co-train the LLM generator and a reward-model-styled detector, embedding the watermarking signature directly within the network weights.

For what concerns the adversarial attacks, we can distinguish between two main strategies, i.e. *untargeted attacks* that include character-level permutations, lexical shuffling, emoji insertions, and automatic paraphrasing or cross-lingual translation workflows, and *targeted attacks* [14] that assume explicit algorithmic awareness to identify the hidden vocabulary partitions via spoofing, watermark stealing through API queries, or structural distillation on watermarked corpora.

In recent years, researchers have increasingly investigated to what extent text watermarking strategies are resilient against translation manipulation. In [15], the authors systematically benchmarked four watermarking strategies under cross-lingual manipulation, evaluating the detection capabilities when translating across diverse and vocabulary-rich languages. Also, in [16] black-box strategies have been evaluated, measuring the resilience of the watermark against four types of post-processing corruptions, highlighting re-translation and structural document polishing as some of the most destructive adversarial environments for token-based systems. Finally, in [17], the authors provide a comprehensive comparison of several representative watermarking strategies across more than 100 languages using 3 main multilingual LLMs. The study shows that while existing methods achieve strong detection performance in their source language, their robustness varies considerably under translation and cross-lingual paraphrasing, with many experiencing substantial degradation in watermark detectability.

All these findings highlight the need to evaluate watermarking techniques in diverse multilingual scenarios and establish multilingual robustness as a key criterion for comparing and benchmarking future watermarking strategies.

Despite these extensive advancements, existing literature exhibits prominent gaps regarding the robustness of text watermarking against cross-lingual adversarial interventions. First, the vast majority of investigations into watermarking vulnerability under adversarial attacks remain mostly centered on the English language, leaving non-English settings understudied. Second, the structural influence of morphological typology (e.g., fusional, agglutinative, or isolating structures) within the intermediate pivot language remains entirely unexplored, masking how different linguistic properties affect watermarks. Finally, it remains unexamined how different text watermarking strategies balance the trade-off between semantic preservation and statistical detectability when subjected to adversarial back-translation attacks according to different structures. Such a final aspect allows us to understand if a worst-case scenario can manifest, i.e. when the watermark has been successfully *removed* (low detectability) and the semantic similarity remains high with respect to a back-translated text maintaining its original meaning.

3. Resilience Evaluation

This study extends the evaluation pipeline designed to quantify the adversarial robustness of text watermarking techniques across diverse Italian LLMs originally introduced in [18]. The overall pipeline, illustrated schematically in Figure 1, comprises four main steps: dataset preparation, watermarked text generation, back-translation attacks and finally evaluation.

While the current and the former papers evaluate the resilience of text watermarking in Italian Large Language Models against back-translation attacks, the current one is focused on pivot languages by morphological typology (fusional, agglutinative, isolating) to analyze the trade-off between detectability and semantic preservation, whereas the previous paper structures pivot languages strictly by resource tiers (high, mid, low) and includes to examine the impact on model perplexity.

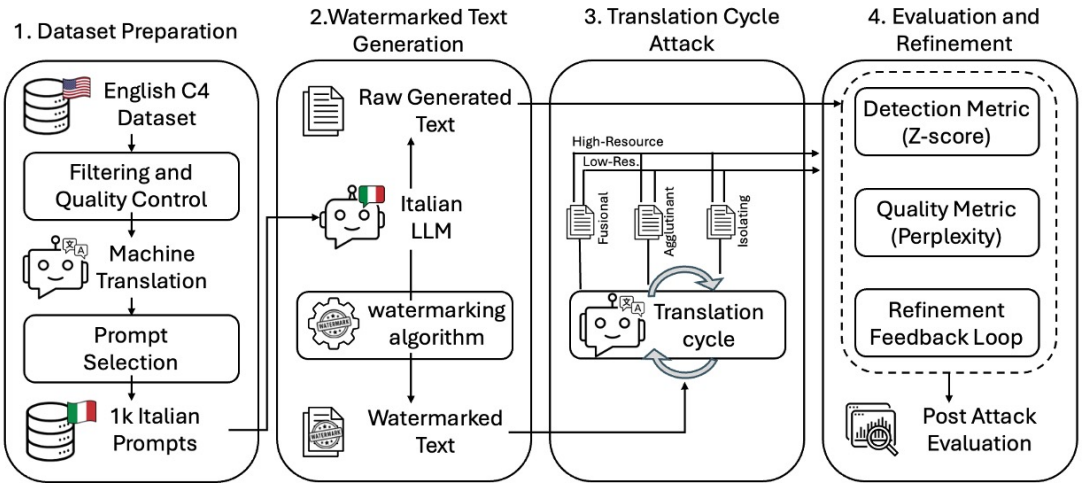


Figure 1: Schematic representation of the proposed evaluation and attack pipeline.

3.1. Dataset Preparation

The dataset preparation phase establishes a curated linguistic baseline for validating watermark resilience in Italian. We develop a systematic filtering pipeline to translate and refine high-quality web-scraped content into standard Italian prompts through three phases:

1. **Filtering and Quality Control**: Starting from the English C4 Dataset [19], we sampled those prompts which lengths are restricted to a minimum of 450 tokens to guarantee adequate generative context while minimizing long-form model drift.
2. **Neural Machine Translation**: The selected prompts are translated into Italian using the specialized X-ALMA model [20]. X-ALMA leverages language-specific modular architectures to maintain semantic coherence, ensuring that the resulting Italian prompts mirror the intent and syntactic complexity of the original ones.
3. **Prompt Selection**: From this processed corpus, a subset of 1,000 prompts has been randomly sampled to measure watermark detectability and semantic similarity with respect to the original text. Moreover, in order to maintain only proper Italian prompts, only the text with at least 99% of the Italian words after the English-to-Italian translation are maintained.

3.2. Watermarked Text Generation

We frame generation as a constrained text completion task where the target model processes the Italian prompt as a prefix and generates a continuation up to $T = 250$ tokens. This sequence length guarantees sufficient statistical signal for detection while matching the optimal generation window of mid-sized foundational models. Watermarks are applied dynamically during the decoding stage by modifying the model's token logit distributions. To thoroughly evaluate watermarking resilience, we select three representative state-of-the-art watermarking strategies:

- **LogitsWatermarker (KGW)** [9]: Serving as the syntactic baseline, KGW uses a pseudo-random function seeded by the hash of the preceding token to split the vocabulary V into a green list G and a red list R. The modified next-token probability distribution $\hat{p}_k^{(t)}$ for a token k at step t is formalized as:

$$\hat{p}_k^{(t)} = \frac{\exp(l_k^{(t)} + \delta \cdot \mathbb{I}_{k \in G})}{\sum_{i \in R} \exp(l_i^{(t)}) + \sum_{i \in G} \exp(l_i^{(t)} + \delta)} \tag{1}$$

  where $l_k^{(t)}$ represents the baseline logit, δ is the constant bias hardness parameter, and $\mathbb{I}_{k \in G}$ is an indicator function mapping token inclusion in the green list.

- **SemaMark** [10]: To bypass the fragility of discrete token hashes, SemaMark projects token history into a continuous semantic space. It aggregates contextual information across a window of m preceding tokens via a weighted pooling of their dense embeddings e_i, mapping them onto a Normalized Embedding Ring (NE-Ring) to compute a semantic seed value P:

$$P(\{e_{i+1:i+m}\}) = \sum_{j=1}^{m} \frac{j + \frac{m}{2}}{w_{\text{sum}}} e_{i+j} \tag{2}$$

  where w_{sum} is a scalar positional normalization factor defined as:

$$w_{\text{sum}} = \sum_{j=1}^{m} \left(j + \frac{m}{2}\right) \tag{3}$$

  This continuous seed dynamically identifies contextually stable green lists based on embedding trajectories.
- **Semantic Invariant Robustness (SIR)** [11]: SIR uses an auxiliary neural encoder E to derive a dense context embedding e_l from the history sequence. A specialized mapping network T subsequently translates this embedding into a continuous watermark logit modulation vector $P_W \in [-1, +1]$:

$$P_{\hat{M}} = P_M + \delta \cdot T(E(t_{0:l-1})) \tag{4}$$

  By relying on dense semantic representations, both SIR and SemaMark hypothesize that the watermark signal will survive surface-level alterations because neural machine translation models preserve underlying semantic invariants during language transfers.

3.3. Adversarial Back-Translation Attack

To rigorously assess watermark resilience against watermarking removal, we implement an untargeted adversarial back-translation pipeline. This paradigm simulates an adversary leveraging machine translation as a semantic-preserving sieve to disrupt localized token statistics while maintaining underlying core semantic text meanings. The attack routes watermarked sequences through a closed translation loop:

$$\text{Italian} \longrightarrow \text{Pivot Language } (\mathcal{L}_{\text{pivot}}) \longrightarrow \text{Italian} \tag{5}$$

Rather than selecting pivot languages arbitrarily, we systematically extended the original framework to comprise $\mathcal{L}_{\text{pivot}}$ along two orthogonal dimensions to isolate the compound effects of cross-lingual structural transformations and data scarcity: morphological typology, i.e. fusional, agglutinative, and isolating, and resource availability, i.e. high- vs. low-resource settings.

- **Fusional Pivots**, employing *German* as the high-resource baseline and *Hindi* as the low-resource counterpart. These languages rely heavily on highly inflected, fusional morphology and complex case-marking systems, forcing extensive sub-word token realignment and structural reorganization during the intermediate encoding and decoding cycles.
- **Agglutinative Pivots**, utilizing *Finnish* and *Uzbek* as high- and low-resource representatives, respectively. Characterized by sequential morpheme stacking onto root forms, these languages trigger severe re-tokenization and morphological expansion, allowing us to evaluate whether the diffusion of semantic seeds across multiple bound morphemes rapidly degrades logit-level signatures.
- **Isolating Pivots** via high-resource *Chinese* and low-resource *Afrikaans*. As isolating languages that strictly eschew inflectional morphology in favor of rigid syntax and standalone aspect particles, they enforce a profound structural "flattening" of Italian's rich verbal morphology, enabling an explicit assessment of how severe shifts in linguistic structure disrupt watermark detectability.

3.4. Evaluation Metrics

In order to understand the watermarking resilience when subject to different pivotal and high/low-resource languages, we extended the original framework with three evaluation metrics:

- **Z-score and True Positive Rate**: The foundational presence of the watermark is measured via the standard Z-score, which quantifies the statistical deviation of greenlist token counts relative to expected baseline distributions:

$$z = \frac{|s|_G - \gamma T}{\sqrt{T\gamma(1-\gamma)}} \tag{6}$$

  where $|s|_G$ denotes total green token occurrences, T indicates total sequence length, and γ represents the structural proportion of the vocabulary assigned to the greenlist. In our analysis, we evaluate empirical detection rates against two distinct operational thresholds: a permissive boundary ($\tau = 2.5$) and a highly conservative, restrictive boundary ($\tau = 4.0$). Since all the considered texts are effectively LLM-generated, we estimated the Detection Rate, i.e., True Positive Rate, according to the two main thresholds. In particular, let $\mathcal{Z} = \{z_1, z_2, \ldots, z_N\}$ be the set of Z-scores computed over a corpus of N watermarked sequences, and let $\tau \in \{2.5, 4.0\}$ denote the detection threshold. The True Positive Rate (TPR_τ) is defined as:

$$\text{TPR} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}\left(z_i \geq \tau\right) \tag{7}$$

  where $\mathbb{I}(\cdot)$ represents the indicator function, returning 1 if the empirical Z-score z_i equals or exceeds the detection threshold τ, and 0 otherwise.
- **Distributional Shift Analysis (KDE and Wasserstein Distance)**: To analyze the macro-level impact of adversarial translation on the watermark signal, we employ Kernel Density Estimation (KDE) to model the shifts in the Z-score distributions. We subsequently compute the Wasserstein distance between the watermarked Italian and back-translated text distributions to quantify the distributional shift induced by each pivot language. More formally, let $\mathcal{Z}_{\text{native}} = \{z_1, \ldots, z_N\}$ and $\mathcal{Z}_{\text{pivot}} = \{z'_1, \ldots, z'_N\}$ denote the sets of Z-scores computed over the native watermarked Italian corpus and the post-back-translation corpus, respectively. The underlying continuous probability density functions, $f(z)$, are estimated as:

$$\hat{f}(z) = \frac{1}{Nh} \sum_{i=1}^{N} K\left(\frac{z - z_i}{h}\right) \tag{8}$$

  where $K(\cdot)$ represents a standard Gaussian kernel function and $h > 0$ denotes the optimal smoothing bandwidth determined via Scott's rule.
  To quantify the vulnerability of the watermark when subject to the back-translation attack, we compute the first Wasserstein distance ($\mathcal{W}_1$) between the empirical cumulative distribution functions F_{native} and F_{pivot}:

$$\mathcal{W}_1\left(F_{\text{native}}, F_{\text{pivot}}\right) = \int_{-\infty}^{+\infty} \left|F_{\text{native}}(z) - F_{\text{pivot}}(z)\right| dz \tag{9}$$

  This metric measures the total transportation cost required to align the Z-score distribution of the back-translated text with the original generation baseline, providing an estimation of the statistical degradation of the watermark signal of each pivot language back-translation.
- **Semantic Preservation (BERTScore)**: To estimate the semantic preservation of a back-traslated text, we compute the BERTScore [21] between the original watermarked Italian text $R = \langle r_1, \ldots, r_m \rangle$ and its back-translated sequence $\hat{C} = \langle \hat{c}_1, \ldots, \hat{c}_n \rangle$. Using a multilingual

contextual encoder, both sequences are mapped to sequence-normalized embeddings $\mathbf{R}$ and $\hat{\mathbf{C}}$, where the token alignment matrix is defined via the pairwise dot product $\mathbf{M}_{ij} = \mathbf{r}_i^\top \hat{\mathbf{c}}_j$. The semantic preservation is quantified by the harmonic mean of the maximum-similarity token mappings, denoted as the F-score:

$$F_{\text{BERT}} = 2 \cdot \frac{P_{\text{BERT}} \times R_{\text{BERT}}}{P_{\text{BERT}} + R_{\text{BERT}}} \tag{10}$$

where $P_{\text{BERT}} = \frac{1}{n} \sum_{\hat{c}_j \in \hat{C}} \max_{r_i \in R} \mathbf{r}_i^\top \hat{\mathbf{c}}_j$ and $R_{\text{BERT}} = \frac{1}{m} \sum_{r_i \in R} \max_{\hat{c}_j \in \hat{C}} \mathbf{r}_i^\top \hat{\mathbf{c}}_j$ correspond to contextual precision and recall, respectively. BERTScores and z-scores will be subsequently used to determine the trade-off between statistical detectability and semantic persistence.

4. Experimental Configuration

4.1. Source Corpus

Prompts are drawn from the standardized AllenAI Clean Crawled Corpus (*C4*) [19], a large-scale collection of natural language web-scraped data. The random sampling of C4 ensures broad domain diversity, allowing us to benchmark watermarking performance across different writing styles and vocabulary distributions while facilitating direct comparison with prior English baselines. The sampled prompts are available at https://github.com/MIND-Lab/WatermarkinItalianLLMs.

4.2. Machine Translation Infrastructure

The adversarial cross-lingual attacks are executed via the X-ALMA engine [20][1]. X-ALMA extends the LLaMA-2 architecture by integrating language-specific sub-modules designed to actively counter cross-lingual interference. By ensuring satisfactory translation quality across both high-resource and low-resource languages, X-ALMA can reduce watermark degradation due to artefacts related to translation errors.

4.3. Italian Large Language Models

Our evaluation suite spans five state-of-the-art Italian-centric LLMs to ensure findings generalize robustly across diverse architectural backbones, vocabulary sizes, and training regimes:

- **Minerva-7B** [22]: A Mistral-based architecture trained from scratch on an evenly balanced, 2.28-trillion-token bilingual corpus of Italian and English.
- **Modello Italia-9B** [23]: An enterprise-focused foundational model optimized intensely for regional nuances, containing over 90% native Italian pre-training data.
- **Miia-7B** [24]: An architecture built around a custom tokenizer optimized heavily for Italian text processing, operating an extended 16K context window.
- **LLaMAntino-3-ANITA-8B** [25]: An Italian-aligned model built on top of the LLaMA-3 architecture, refined using sequential Supervised Fine-Tuning (SFT) and Direct Preference Optimization (DPO).
- **Velvet-14B** [26]: A large-scale model featuring a 127K vocabulary and an expansive 131K context window, engineered specifically to preserve long-range text coherence across industrial domains.

A comparative breakdown of model parameters and properties is provided in Table 1.

[1]X-ALMA has been selected among three NMT architectures, i.e., NLLB-54B, X-ALMA-13B and LLaMAX3-8B-Alpaca, because of its dominance in terms of xComet, MetricX and CometWiki

Table 1
Technical specifications of the evaluated Italian language model suite.

Model Identifier	Parameters	Context Window	Hugging Face Repository Reference
`Minerva-7B` [22]	7 Billion	4,096	`sapienzanlp/Minerva-7B-instruct-v1.0`
`Modello Italia-9B` [23]	9 Billion	4,096	`sapienzanlp/modello-italia-9b-bf16`
`Miia-7B` [24]	7 Billion	16,384	`Fastweb/FastwebMIIA-7B`
`Velvet-14B` [26]	14 Billion	131,072	`Almawave/Velvet-14B`
`LLaMAntino-3-8B` [25]	8 Billion	8,192	`swap-uniba/LLaMAntino-3-ANITA-8B-Inst-DPO-ITA`

5. Experimental Results

5.1. Effect of back-translation

Figure 2 illustrates an example of the kernel density estimates (KDEs) of the z-score distributions related to the Native (Italian) watermarked text and the one subject to the back-translation manipulation, related to the LLamantino model under the KGW watermarker[2]. Across all language models, the z-score distribution of the back-translated texts is consistently shifted toward lower values compared with the corresponding distribution obtained from the Native (Italian) watermarked texts. As a result, a substantial portion of the attacked samples falls below the detection thresholds τ=2.5 and τ=4. This behavior indicates that back translation significantly weakens the statistical watermark signal, making the attacked texts considerably more difficult to identify as watermarked. The effect is particularly evident for the more conservative threshold (τ=4), where only a small fraction of the attacked distribution remains above the decision boundary. Overall, these results demonstrate that back-translation attacks substantially reduce watermark detectability under both detection thresholds, highlighting the vulnerability of current watermarking methods to translation based transformations.

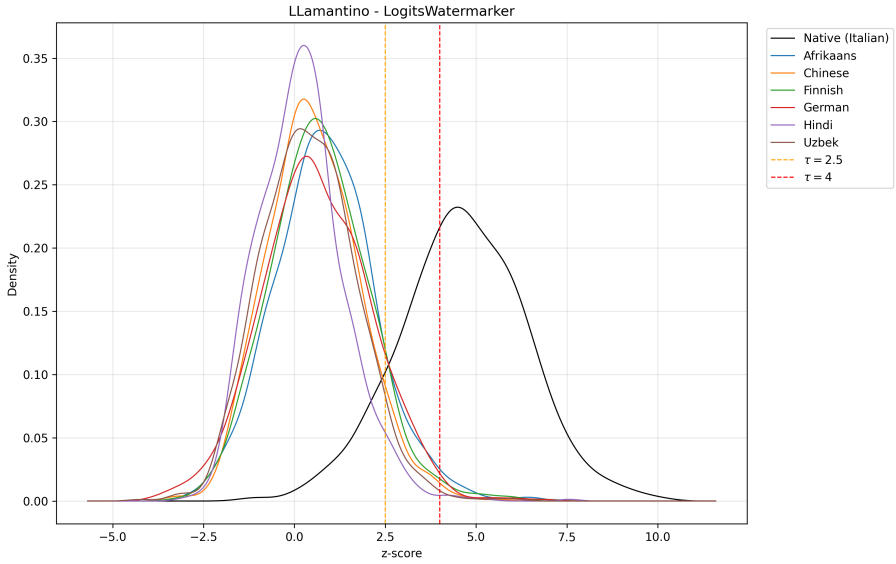


Figure 2: Kernel Density Estimations of KGW on LLamantino

In order to quantify the effect of the attacks, Table 2 reports the Wasserstein distance between the z-score distribution related to the native (Italian) watermarking and the z-score distribution related to the back-translation languages. Overall, the results reveal consistent differences across linguistic typologies. Among the fusional languages (German and Hindi), Hindi consistently produces larger Wasserstein distances than German across all language models and watermarking methods. The same trend is observed for SemaMark and SIR, indicating that back translation through Hindi induces

[2]Similar results have been obtained for the other LLMs and watermarking strategies

larger shifts in the z-score distribution than back translation through German. For the agglutinative languages (Finnish and Uzbek), Uzbek generally exhibits larger Wasserstein distances than Finnish. This behavior is consistent across all models and is particularly pronounced for KGW and SemaMark. The isolating languages (Chinese and Afrikaans) show more heterogeneous behavior. Chinese consistently produces intermediate Wasserstein distances, whereas Afrikaans yields the smallest values in nearly all experimental settings. Consequently, back translation through Afrikaans alters the watermark score distribution considerably less than the other pivot languages considered. Overall, the experiments suggest that the effect of back translation varies substantially across pivot languages, with the largest distributional shifts observed for Hindi and Uzbek, and the smallest for Afrikaans and German.

Focusing on the pivot languages according to resource availability reveals a less clear trend. Among the high-resource languages (German, Finnish, and Chinese), the Wasserstein distances remain relatively moderate, with German consistently producing the smallest values within the group and Chinese the largest. Finnish generally lies between the two. The low-resource languages (Hindi, Uzbek, and Afrikaans) exhibit considerably higher variability. Hindi and Uzbek consistently produce the largest Wasserstein distances across language models and watermarking methods, whereas Afrikaans behaves similarly to German, producing among the smallest distances overall. This observation suggests that resource availability alone does not adequately explain the magnitude of the distributional shift introduced by back translation.

	Llamantino					
	Afrikaans	**Chinese**	**Finnish**	**German**	**Hindi**	**Uzbek**
KGW	3.781	4.102	3.921	4.012	4.429	4.229
SemaMark	3.841	4.338	3.843	3.637	5.273	4.983
SIR	2.690	3.050	2.680	2.360	4.953	3.698
	Minerva					
	Afrikaans	**Chinese**	**Finnish**	**German**	**Hindi**	**Uzbek**
KGW	9.308	9.568	9.612	8.846	10.941	11.055
SemaMark	7.291	7.808	7.003	6.366	9.973	9.836
SIR	5.144	5.694	5.314	4.382	7.020	7.031
	Velvet					
	Afrikaans	**Chinese**	**Finnish**	**German**	**Hindi**	**Uzbek**
KGW	8.148	7.779	7.991	7.209	9.353	9.27
SemaMark	4.999	5.721	5.532	4.839	6.319	7.278
SIR	4.363	5.054	4.507	3.831	5.944	5.930
	MiiA					
	Afrikaans	**Chinese**	**Finnish**	**German**	**Hindi**	**Uzbek**
KGW	8.445	8.315	8.433	7.646	9.881	9.851
SemaMark	6.553	10.722	10.827	5.616	9.542	10.851
SIR	3.870	4.353	4.110	3.366	5.809	5.916
	Modello Italia					
	Afrikaans	**Chinese**	**Finnish**	**German**	**Hindi**	**Uzbek**
KGW	9.015	8.983	9.116	8.205	10.631	10.57
SemaMark	6.190	7.129	6.235	5.185	8.710	8.869
SIR	3.686	4.394	3.815	3.199	6.016	5.950

Table 2
Wasserstein distance between the Native (Italian) watermaked z-score distribution and the z-score distribution related to each pivot language adopted for back-translation.

We finally report in Figure 3, the results related to True Positive Rates, highlighting how the model behaviour changes across languages and watermarking strategies (considering τ=2.5).

Several consistent trends emerge across the evaluated language models. First, the choice of pivot language has a substantial impact on watermark detectability. For most models and watermarking strategies, German yields the highest TPR values, indicating that back translation through German

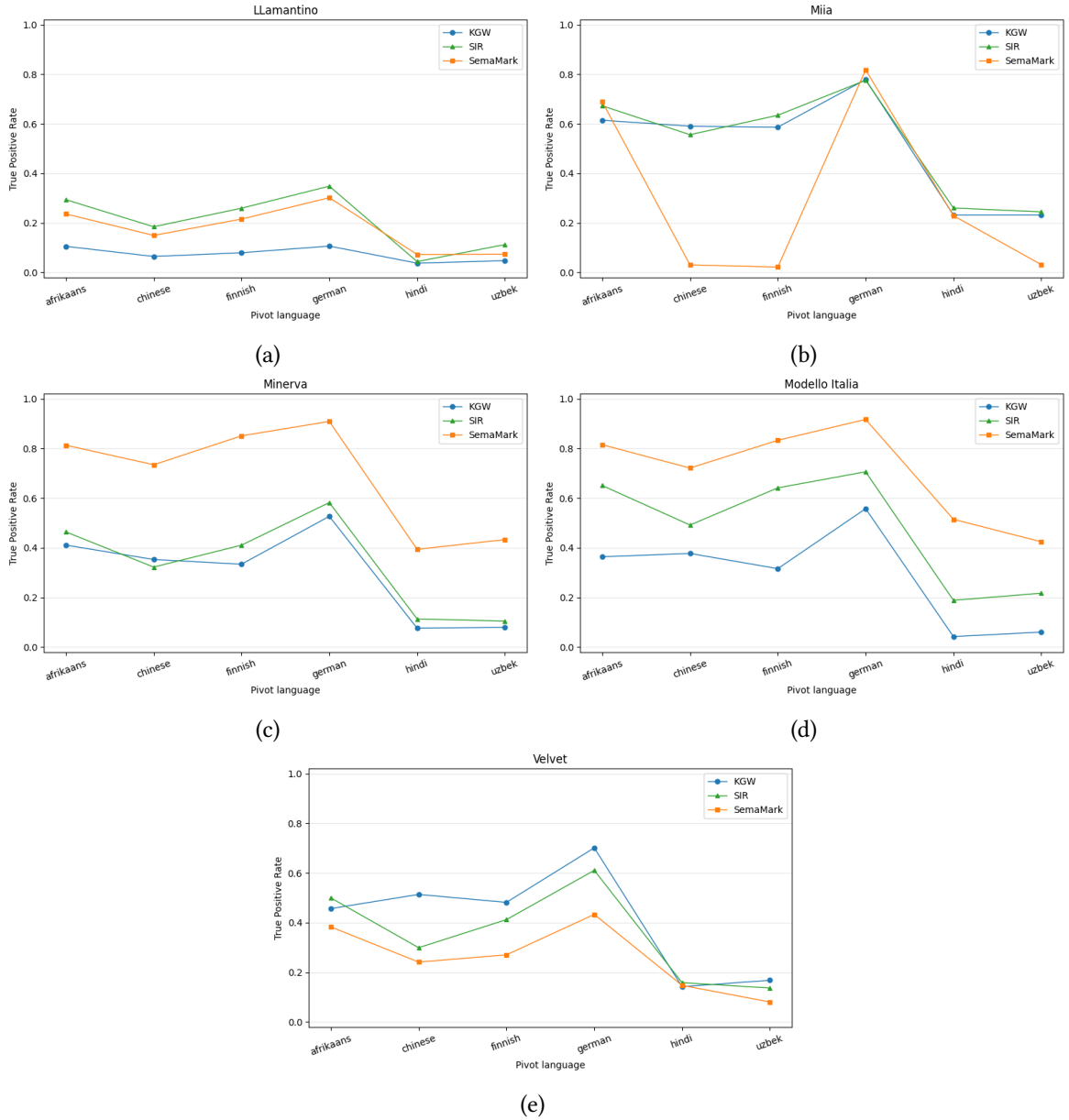


Figure 3: Comparison of True Positive Rate (TPR) across five language models subject to three watermarking strategies (KGW, SIR, SemaMark) evaluated on six pivot languages.

preserves a larger fraction of the watermark signal. In contrast, Hindi and Uzbek systematically lead to the lowest TPR values, often causing a dramatic reduction in detection performance. This observation is consistent with the distributional analysis presented in the previous section, where these languages were associated with the largest shifts in the z-score distributions. Chinese and Finnish generally produce intermediate results, while Afrikaans often preserves watermark detectability better than the other non-European pivot languages.

The robustness of the watermarking methods also varies considerably across models. SemaMark achieves the highest TPR in Minerva and Modello Italia, maintaining detection rates above 0.7 for several pivot languages and remaining comparatively resilient even under the more challenging Hindi and Uzbek attacks. In contrast, KGW exhibits a stronger sensitivity to back translation, particularly for Llamantino, where the TPR remains below 0.12 across all languages, and for Modello Italia, where detection rates drop to nearly zero for Hindi and Uzbek. SIR generally occupies an intermediate position, although it becomes the most robust strategy for Llamantino and performs comparably to KGW in MiiA. The results further highlight substantial differences among language models. MiiA consistently achieves

the highest detection rates, with TPR values frequently exceeding 0.6 and reaching approximately 0.8 for German, whereas Llamantino appears particularly vulnerable to back-translation attacks regardless of the watermarking strategy employed.

Taken together, these findings demonstrate that the effectiveness of watermark detection depends not only on the watermarking algorithm but also on the interaction between the underlying language model and the pivot language used during back-translation. The pronounced performance degradation observed for Hindi and Uzbek suggests that certain linguistic transformations introduced during translation can substantially weaken the watermark signal, reducing the reliability of watermark detection even when the original watermarked text is detected with high confidence.

5.2. Vulnerability of the Watermarking Strategies

To measure the extent of the vulnerability of each watermarking strategy, we analyze the empirical *Vulnerability*, defined as the percentage of generated sequences that have a (BERTscore ≥ 0.7) and the z-score ($z \leq 4.0$) after a back-translation manipulation. Analysing the watermarking strategies f across models reveals systematic differences in terms of vulnerabilities. In Figure 4, as a representative example, we present the results for KGW under back-translation via German, which yielded the least degradation among the languages evaluated in Section 5.1. The chart reports the z-scores on the X-axis

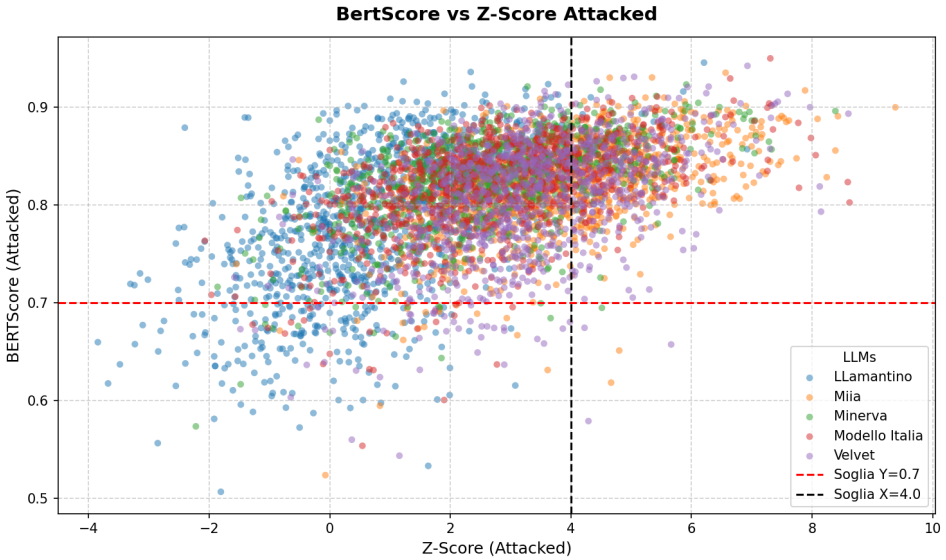


Figure 4: KGW vulnerability through the German back-translation attack.

and the BertScore on the Y-axis. The top-right area delimited by the dashed lines, comprising instances having a BertScore greater than 0.7 and a z-score less than 4.0, represent the most problematic cases. In fact, such instances maintain a semantic similarity with the original Italian text, but with reduced detectability as generated text. This example clearly shows, using the less impactful back-translation (German), the vulnerability of such a watermarking strategy across all the models.

In order to summarize such vulnerability across LLMs and watermarking strategies, we report a summary in Table 3, where the less impactful language has been used as pivot language (German).

The experimental results highlight three primary remarkable considerations:

1. **KGW:** The LogitsWatermaker exhibits extreme vulnerability. For instance, `LLaMAntino-3-8B` undergoes a near-complete erasure of its watermarking signature, resulting in a vulnerability rate of 98.3%. Similarly, models like `Minerva-7B` (80.9%), `Modello Italia-9B` (78.3%), and `Velvet-14B` (70.7%) are characterized by a high vulnerability as well.
2. **SemaMark:** The continuous embedding alignment of SemaMark achieves the lowest vulnerability rates. The framework reaches optimal efficiency when paired with `Modello Italia-9B` and

Table 3
Vulnerability (%) under adversarial back-translation manipulation across Italian LLMs.

Model	KGW	SemaMark	SIR
`LLamantino-3-8B`	98.3%	84.4%	90.4%
`FastwebMIIA-7B`	55.7%	35.4%	48.2%
`Minerva-7B`	80.9%	22.2%	64.5%
`Modello Italia-9B`	78.3%	20.8%	49.3%
`Velvet-14B`	70.7%	79.9%	67.1%

`Minerva-7B`, suppressing text compromise down to of 20.8% and 22.2%, respectively. Although the vulnerability rate decreases, the number of instances that represent the most problematic ones to detect remains remarkably high.

3. **SIR:** The SIR watermarking strategy preserves high semantic similarity (BERTscore ≥ 0.80), but it demonstrates a higher vulnerability across the entire model suite compared to SemaMark. It surfaces a high vulnerability of 90.4% on `LLaMAntino-3-8B` and in the remaining cases, the vulnerability appears to be more $\sim 50\%$. Even within highly optimized foundational models like `Modello Italia-9B` and `Velvet-14B`, SIR shows to be vulnerable.

This consistent behaviour in vulnerability suggests that all the Italian LLMs and the considered watermarking strategies are strongly affected by the back-translation attacks. This highlights a key limitation in current watermarking paradigms, pointing to cross-lingual stability as a crucial direction for future research.

6. Conclusions and Future Work

In this work, we systematically investigated the adversarial robustness of generation-time text watermarking strategies within Italian Large Language Models when subjected to cross-lingual back-translation attacks. By benchmarking three representative watermarking paradigms—LogitsWatermark (KGW), SemaMark, and SIR across five Italian-centric LLMs, our findings reveal that back-translation serves as a highly effective semantic-preserving sieve that significantly degrades statistical watermark signals. Our evaluation across morphological typologies and resource availability demonstrates that pivot languages with complex inflectional or agglutinative structures, such as Hindi and Uzbek, induce the most severe z-score distributional shifts. Furthermore, analyzing the trade-off between semantic preservation and statistical detectability reveals widespread vulnerability. Token-level logit-biasing schemes experience near-total detection collapse (up to 98.3% compromise), while continuous semantic embedding approaches—though substantially more resilient—still exhibit notable vulnerability (ranging down to 20.8%). Overall, these results demonstrate that current state-of-the-art watermarking frameworks are extremely vulnerable when attacked adopting multilingual manipulations.

Building upon our empirical insights, several promising directions emerge for future research. A first research direction relates the development of novel decoding-time watermarking strategies that explicitly leverage cross-lingual semantic invariants to ensure watermark persistence regardless of translation loops. An additional aspect, relates to the adoption of morphology-aware watermark generation, by the design of adaptive hashing and semantic seeding mechanisms that account for morphological transformations, preventing the watermark removal when text transitions between highly inflected, agglutinative, or isolating structures. An additional research direction relates the evaluation beyond zero-bit detection toward multi-bit payloads, assessing whether structured payload signatures offer higher resilience against back-translation attacks. Finally, targeted adversarial strategies need to be explored to define defense mechanisms, such as training robust multi-lingual watermark detectors or co-training via reinforcement learning, to make models more robust against NMT-driven manipulations.

Declaration on Generative AI

During the preparation of this work, the author(s) used Gemini in order to rephrase sentences to improve clarity, conciseness, or style. After using this service, the author(s) reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] A. Bondielli, G. Bonetta, E. Leonardelli, I. Siragusa, Preface to the Ninth Workshop on Natural Language for Artificial Intelligence (NL4AI), in: Proceedings of the Ninth Workshop on Natural Language for Artificial Intelligence (NL4AI 2026) co-located with 24th International Conference of the Italian Association for Artificial Intelligence (AI*IA 2026), 2026.

[2] A. Liu, L. Pan, Y. Lu, J. Li, X. Hu, X. Zhang, L. Wen, I. King, H. Xiong, P. Yu, A survey of text watermarking in the era of large language models, ACM Computing Surveys 57 (2024) 1–36.

[3] Z. He, B. Zhou, H. Hao, A. Liu, X. Wang, Z. Tu, Z. Zhang, R. Wang, Can watermarks survive translation? on the cross-lingual consistency of text watermark for large language models, in: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2024, pp. 4115–4129.

[4] J. T. Brassil, S. Low, N. F. Maxemchuk, L. O'Gorman, Electronic marking and identification techniques to discourage document copying, IEEE Journal on Selected Areas in Communications 13 (1995) 1495–1504.

[5] E. German, S. Antebi, E. Habler, A. Shabtai, Y. Elovici, Leximark: Robust watermarking via lexical substitutions to enhance membership verification of an llm's textual training data, arXiv preprint arXiv:2506.14474 (2025).

[6] A. Nemecek, Y. Jiang, E. Ayday, Topic-based watermarks for large language models, in: Findings of the Association for Computational Linguistics: ACL 2026, 2026, pp. 24372–24402.

[7] S. Park, H. Park, H. An, Y.-S. Han, A linguistics-aware llm watermarking via syntactic predictability, in: Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2026, pp. 45629–45647.

[8] J. Kim, S. Park, Y.-S. Han, Marking code without breaking it: Code watermarking for detecting llm-generated code, in: Findings of the Association for Computational Linguistics: EACL 2026, 2026, pp. 3990–4002.

[9] J. Kirchenbauer, J. Geiping, Y. Wen, J. Katz, I. Miers, T. Goldstein, A watermark for large language models, in: International conference on machine learning, PMLR, 2023, pp. 17061–17084.

[10] J. Ren, H. Xu, Y. Liu, Y. Cui, S. Wang, D. Yin, J. Tang, A robust semantics-based watermark for large language model against paraphrasing, in: Findings of the Association for Computational Linguistics: NAACL 2024, 2024, pp. 613–625.

[11] A. Liu, L. Pan, X. Hu, S. Meng, L. Wen, A semantic invariant robust watermark for large language models, in: International Conference on Learning Representations, volume 2024, 2024, pp. 6499–6519.

[12] R. Kuditipudi, J. Thickstun, T. Hashimoto, P. Liang, Robust distortion-free watermarks for language models, Transactions on Machine Learning Research (????).

[13] A. Hou, J. Zhang, Y. Wang, D. Khashabi, T. He, k-semstamp: A clustering-based semantic watermark for detection of machine-generated text, in: Findings of the Association for Computational Linguistics: ACL 2024, 2024, pp. 1706–1715.

[14] Y. Cheng, H. Guo, Y. Li, L. Sigal, Revealing weaknesses in text watermarking through self-information rewrite attacks, in: Forty-second International Conference on Machine Learning, 2025. URL: https://openreview.net/forum?id=fE3kgW7kMp.

[15] M. Al Ghanim, J. Xue, R. P. Hastuti, M. Zheng, Y. Solihin, Q. Lou, Evaluating the robustness and accuracy of text watermarking under real-world cross-lingual manipulations, in: C. Christodoulopoulos, T. Chakraborty, C. Rose, V. Peng (Eds.), Findings of the Association for Computational Linguistics:

EMNLP 2025, Association for Computational Linguistics, Suzhou, China, 2025, pp. 7396–7416. URL: https://aclanthology.org/2025.findings-emnlp.390/. doi:10.18653/v1/2025.findings-emnlp.390.

[16] J. Hao, J. Qiang, Y. Zhu, Y. Li, Y. Yuan, X. Ouyang, Post-hoc watermarking for robust detection in text generated by large language models, in: O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, S. Schockaert (Eds.), Proceedings of the 31st International Conference on Computational Linguistics, Association for Computational Linguistics, Abu Dhabi, UAE, 2025, pp. 5430–5442.

[17] A. Mohamed, M. Gubri, Is multilingual LLM watermarking truly multilingual? scaling robustness to 100+ languages via back-translation, in: Trustworthy AI for Good (AI4GOOD) Workshop @ ICML 2026, 2026. URL: https://openreview.net/forum?id=IbUbUbcZ2b.

[18] D. Altieri, M. Pasotti, E. Opreni, A. Pietrasanta, S. Claudio, M. Pasqualini, G. Rizzi, E. Fersini, Lost in translation? watermark resilience and perplexity degradation in italian llms across resource tiers, in: Proceedings of the Twelfth Italian Conference on Computational Linguistics (CLiC-it 2026), 2026.

[19] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, Journal of machine learning research 21 (2020) 1–67.

[20] H. Xu, K. Murray, P. Koehn, H. Hoang, A. Eriguchi, H. Khayrallah, X-ALMA: Plug & play modules and adaptive rejection for quality translation at scale, in: The Thirteenth International Conference on Learning Representations, 2025.

[21] T. Zhang*, V. Kishore*, F. Wu*, K. Q. Weinberger, Y. Artzi, Bertscore: Evaluating text generation with bert, in: International Conference on Learning Representations, 2020. URL: https://openreview.net/forum?id=SkeHuCVFDr.

[22] R. Orlando, L. Moroni, P.-L. Huguet Cabot, S. Conia, E. Barba, S. Orlandini, G. Fiameni, R. Navigli, Minerva LLMs: The first family of large language models trained from scratch on Italian data, in: F. Dell'Orletta, A. Lenci, S. Montemagni, R. Sprugnoli (Eds.), Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024), CEUR Workshop Proceedings, Pisa, Italy, 2024, pp. 707–719.

[23] T. iGenius Team, Italia-9b-instruct-v0.1, 2024. URL: https://huggingface.co/iGeniusAI/Italia-9B-Instruct-v0.1.

[24] Fastweb, Fastwebmiia-7b, 2025. URL: https://huggingface.co/Fastweb/FastwebMIIA-7B.

[25] M. Polignano, P. Basile, G. Semeraro, Advanced natural-based interaction for the italian language: Llamantino-3-anita, Scientific Reports (2026).

[26] Almawave, Velvet-14b, 2025. URL: https://huggingface.co/Almawave/Velvet-14B.