Authorship Attribution in Multilingual Machine-Generated Texts

Lucio La Cava[1], **Dominik Macko**[2], **Robert Moro**[2], **Ivan Srba**[2], **Andrea Tagarelli**[1]
[1]DIMES Department, University of Calabria, Italy
[2]Kempelen Institute of Intelligent Technologies, Slovakia
{lucio.lacava, tagarelli}@dimes.unical.it
{dominik.macko, robert.moro, ivan.srba}@kinit.sk

Abstract

As Large Language Models (LLMs) have reached human-like fluency and coherence, distinguishing machine-generated text (MGT) from human-written content becomes increasingly difficult. While early efforts in MGT detection have focused on binary classification, the growing landscape and diversity of LLMs require a more fine-grained yet challenging *authorship attribution* (AA), i.e., being able to identify the precise generator (LLM or human) behind a text. However, AA remains nowadays confined to a monolingual setting, with English being the most investigated one, overlooking the multilingual nature and usage of modern LLMs. In this work, we introduce the problem of *Multilingual Authorship Attribution*, which involves attributing texts to human or multiple LLM generators across diverse languages. Focusing on 18 languages—covering multiple families and writing scripts—and 8 generators (7 LLMs and the human-authored class), we investigate the multilingual suitability of monolingual AA methods in terms of their cross-lingual transferability, and the impact of generators on attribution performance. Our results reveal that while certain monolingual AA methods can be adapted to multilingual settings, significant limitations and challenges remain, particularly in transferring across diverse language families, underscoring the complexity of multilingual AA and the need for more robust approaches to better match real-world scenarios.

1 Introduction

Large Language Models (LLMs) have nowadays reached a level of fluency and coherence that enables them to produce human-like text that is no longer distinguishable from that written by humans (Jakesch et al., 2023). While these advancements pave the way for new opportunities in communication, creativity, and productivity (Bubeck et al., 2023), they also raise critical risks in our society about transparency, accountability, and misuse. In particular, the inability to effectively determine whether a text has been generated by humans or machines leaves many open risks for our society, such as misinformation (Chen and Shu, 2024), disinformation (Zugecova et al., 2025), and copyright infringement (Liu et al., 2024).

Early attempts to address the above challenge focused on binary *machine-generated text* (MGT) detection, i.e., automated approaches for distinguishing (AI-based) synthetically generated text from human-written text. However, while effective in many contexts, binary detection has faced a strong limitation: the inability to account for a growing diversity of LLM generators. Indeed, as the number of released LLMs continues to expand day by day, so does the need for fine-grained *authorship attribution* (AA): not just identifying that a text is machine-generated, but also determining which model produced it (Uchendu et al., 2020).

Despite this, existing attempts to perform authorship attribution remain confined to a monolingual setting—with English being the most prominent. This represents a critical blind spot, since modern LLMs are increasingly multilingual, trained to generate content in a broad range of languages, and used in diverse linguistic and cultural contexts.

To address this gap, in this work, we define and investigate the problem of *multilingual authorship attribution*, i.e., attributing texts to the corresponding generators (being they LLMs or humans), across multiple languages and writing scripts. In particular, our study aims to evaluate the multilingual suitability and cross-lingual generalizability of existing AA approaches in this challenging setting, through the following research questions:

RQ1 — *How effectively can existing authorship attribution methods handle multilingual machine-generated text (ML-MGT)?*

RQ2 — *To what extent can authorship attribution approaches for ML-MGT transfer across different languages and language families?*

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 45136–45152

RQ3 — *How does the choice of generator model influence the multilingual suitability and cross-lingual generalizability of authorship attribution methods?*

Contributions. By answering these research questions, our contributions in this work are as follows:

- We introduce and formally define the problems of ML-MGT and Cross-lingual Machine-generated Text (CL-MGT) Authorship Attribution, which handle attributing texts to their machine/human generators across multiple languages and families.
- We evaluate the suitability of existing monolingual authorship attribution methods to the multilingual setting, analyzing how well current monolingual approaches perform in this more challenging scenario, covering 18 languages and 8 different generators.
- We investigate the cross-lingual transferability of authorship attribution methods, assessing their robustness when applied to previously unseen languages.

Our findings suggest that while most existing authorship attribution methods can be extended to the multilingual setting, with varying degrees of efficacy, several challenges persist. Indeed, current authorship attribution methods struggle to generalize across dissimilar language families or writing scripts, as performances are heavily affected by the linguistic properties of the target languages and the identity of the generators. These points underscore the challenges introduced by our newly defined ML-MGT and CL-MGT problems and highlight the pressing need to develop more robust, language-agnostic attribution methods capable of handling the linguistic and stylistic diversity present in real-world multilingual scenarios.

2 Related Work

The human-like text generation capabilities achieved by LLMs in recent years have blurred the distinction between human-authored and machine-generated texts, intensifying the need for reliable detection methods (Jawahar et al., 2020; Crothers et al., 2023; Tang et al., 2024; Wu et al., 2025).

MGT Detection. In response to this challenge, we witnessed a surge in the development of detection methods. These include statistical learning approaches such as probabilistic modeling (Mitchell et al., 2023; Bao et al., 2023; Wang et al., 2023; Miao et al., 2024), log-rank (Su et al., 2023) and perplexity-based methods (Vasilatos et al., 2023), and stylistic or discourse-based approaches (Kim et al., 2024; Gehrmann et al., 2019; Tulchinskii et al., 2023; Venkatraman et al., 2024). Also, watermarking techniques were developed to embed signals in generated texts that remain invisible to humans but are algorithmically detectable (Kirchenbauer et al., 2023; Yoo et al., 2023; Xu et al., 2024) for post-hoc detection. More recently, learning-based methods have gained traction, including deep neural classifiers (Ippolito et al., 2020; Verma et al., 2024), contrastive learning frameworks (Bhattacharjee et al., 2023, 2024), the use of ChatGPT itself as a detector (Bhattacharjee and Liu, 2024), and hybrid approaches incorporating topological features (Uchendu et al., 2023b).

MGT Authorship Attribution. As the diversity of generative models continues to grow, researchers have begun shifting their focus from mere detection to the more ambitious task of *authorship attribution*. This task requires identifying which specific model produced a given text (Uchendu et al., 2020), with important implications for accountability, provenance tracking, and mitigation of misuse (Huang et al., 2025; Uchendu et al., 2023a).

Early works explored the possibility of attributing texts to generators through statistical signals (Solaiman et al., 2019; Gehrmann et al., 2019), but fell short in performance as shown in (La Cava and Tagarelli, 2025). More recent approaches adopt deep learning and contrastive learning strategies, showing stronger results in controlled settings (Guo et al., 2024; La Cava et al., 2024; He et al., 2024). Nevertheless, the body of work on attribution is relatively limited compared to detection.

Multilingual MGT Authorship Attribution. Despite growing attention to attribution, the entire line of research remains fundamentally monolingual, with a predominant focus on English (Wang et al., 2024a; La Cava et al., 2024). A handful of studies have extended to Russian (Shamardina et al., 2022) and Spanish (Sarvazyan et al., 2023), but a systematic investigation of multilingual attribution and the related impact of languages remains underexplored.

This lack motivates our work, and the investigation of multilingual authorship attribution and cross-lingual transferability of attribution methods, as formalized next.

3 Problem Statement

Let us denote with $\mathcal{L}$ a set of *languages* and with $\mathcal{M}$ a set of *machine generators*, i.e., LLMs producing MGTs. *Authorship attribution of multilingual machine-generated text* (ML-MGT) can be formulated as a multi-class classification problem, defined as follows.

Problem 1 (ML-MGT) *We are given a set of texts $\mathcal{X} = \mathcal{X}_h \cup \mathcal{X}_m$, consisting of two subsets: $\mathcal{X}_h$, which contains human-written texts, and $\mathcal{X}_m$, which contains machine-generated texts (MGTs) from all models in $\mathcal{M}$. Each text in $\mathcal{X}_h$ and $\mathcal{X}_m$ is written in a language from the set $\mathcal{L}$. Accordingly, we express these subsets as $\mathcal{X}_h = \bigcup_{\ell \in \mathcal{L}} \mathcal{X}_{h,\ell}$ and $\mathcal{X}_m = \bigcup_{\ell \in \mathcal{L}} \mathcal{X}_{m,\ell}$, where $\mathcal{X}_{h,\ell}$ and $\mathcal{X}_{m,\ell}$ denote the human-written and MGTs in language ℓ, respectively.*

If we denote with y_h the 'HUMAN' class label and with $\mathcal{Y}_m = \{y_j\}_{j=1}^{|\mathcal{M}|}$ the set of 'MACHINE' class labels, the task is to recognize the author of a given text choosing among the human (y_h) and the machine generators in $\mathcal{M}$, i.e., to learn a mapping function $f : \widehat{\mathcal{X}} \mapsto \mathcal{Y} = \{y_h\} \cup \mathcal{Y}_m$, with $\widehat{\mathcal{X}} \subseteq \mathcal{X}$.

In Problem 1, the choice of $\widehat{\mathcal{X}} = \widehat{\mathcal{X}}_h \cup \widehat{\mathcal{X}}_m$ relies on the definition of a *language-selection strategy* $g(\cdot)$ such that, for any $L', L'' \subseteq \mathcal{L}$, $\widehat{\mathcal{X}}_h = g(\mathcal{X}_h, L')$ and $\widehat{\mathcal{X}}_m = g(\mathcal{X}_m, L'')$ are the subsets of $\mathcal{X}_h$, resp. $\mathcal{X}_m$, which select the texts written in any language in L', resp. L''. Unless otherwise specified, we hereinafter assume that $L' = L''$, which implies that *human-written texts and MGTs are provided in the same languages and aligned in a pairwise fashion.*

Problem 2 (CL-MGT) *Let $\mathcal{L}_{train} \subseteq \mathcal{L}$ be the set of languages used for training f, and $\mathcal{L}_{test} \subseteq \mathcal{L}$ be the set of test languages. Problem 1 reduces to an instance of* cross-lingual transferability *if $\mathcal{L}_{train} \subset \mathcal{L}_{test}$.*

The cross-lingual transferability problem aims to evaluate how well a model trained on a set of *source* languages can generalize to *target* languages that were not seen during training. If the test set includes additional languages not seen during training, then the model must rely on its ability to transfer knowledge across languages.

4 Data and Generator Models

To conduct our study, we resorted to the MULTITUDE (v3) dataset (Macko et al., 2025b,a). It contains LLM-generated and human-written news articles, where the latter come from the *MassiveSum* collection (Varab and Schluter, 2021). The machine-generated counterparts are generated by seven LLMs prompted with the original headlines of the articles. These LLMs cover a representative body of open and commercially licensed families of models, spanning various model sizes, architectures, and pre-training strategies, namely *Mistral-7B-Instruct-v0.2*, *OPT-IML-Max-30B*, *v5-Eagle-7B-HF*, *Vicuna-13B*, *Llama-2-70B-Chat-HF*, *Aya-101*, and *GPT-3.5-Turbo-0125*.

Family	Language	Code	Train	Test
Germanic	Dutch	nl	7958	2386
	English	en	7954	2384
	German	de	7951	2388
Hellenic	Greek	el	7944	2384
Semitic	Arabic	ar	7975	2392
Sino-Tibetan	Chinese	zh	7926	2383
Slavic-Cyrillic	Bulgarian	bg	7954	2386
	Ukrainian	uk	7939	2385
	Russian	ru	7945	2382
Slavic-Latin	Croatian	hr	7951	2384
	Czech	cs	7962	2389
	Polish	pl	7946	2383
	Slovak	sk	7946	2385
	Slovenian	sl	7947	2386
Romanic	Portuguese	pt	7956	2388
	Romanian	ro	7949	2386
	Spanish	es	7947	2387
Uralic	Hungarian	hu	7964	2385
Total	–	–	**143,114**	**42,943**

Table 1: Per-language sample counts for train/test splits of the selected data from the MULTITUDE dataset.

Our choice over other existing multilingual MGT datasets (cf. Appendix B), such as M4GT-Bench (Wang et al., 2024b) or RAID (Dugan et al., 2024), was driven by the consistent set of generators, text-generation settings, and domains for each language, enabling focus on unbiased cross-lingual transferability aspects.

Among the 21 languages available in MULTITUDE, we focused on the 18 languages that (i) provide fully balanced coverage across language-generator combinations, to avoid skewed or under-represented distributions that could bias evaluation, and (ii) contain at least 95% of the target number of samples—1,000 per generator for training and 300 per generator for testing—for a robust and fair comparison across languages and models. Table 1 provides statistics on the train-test splits. Each value

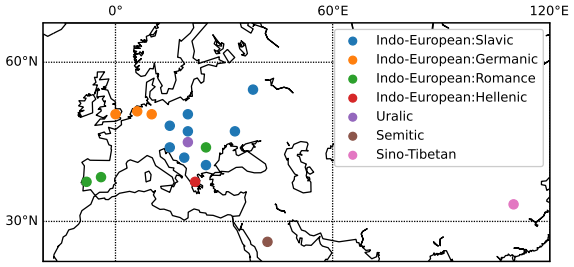

Figure 1: Language coverage for our multilingual AA.

reflects a uniform distribution across all classes, with 1/8 of the samples assigned to human-written texts, and the remainder evenly distributed across the LLM generators.

Language analysis. As shown in Fig. 1, our selected data covers eight language families, namely Indo-European—organized into Germanic, Romanic, Slavic-Latin, Slavic-Cyrillic, and Hellenic—Uralic, Semitic, and Sino-Tibetan. This also corresponds to five writing scripts (12×Latin, 3×Cyrillic, 1×Arabic, 1×Hanzi, and 1×Greek). Thus, the selected language composition enables various combinations of investigations and in-depth insights regarding multilingual and cross-lingual characteristics of AA methods.

5 Detection Methods

In this section, we present the methods selected for evaluation in a multilingual setting. These were chosen based on their strong performance in recent works (Sarvazyan et al., 2023; He et al., 2024; Wang et al., 2024a; La Cava and Tagarelli, 2025). However, all of them required adaptation to suit the specific demands of our target attribution problems, i.e., ML-MGT and CL-MGT. Next, we detail the adaptation process for each method.

5.1 Statistical Approaches

Individual statistical approaches. We consider two zero-shot binary detectors to extract statistical features from texts, i.e., Fast-DetectGPT (Bao et al., 2023) with mGPT-13B[1] by Shliazhko et al., 2024 as both the reference and sampling model, and Binoculars (Hans et al., 2024) with Falcon-7B by Almazrouei et al., 2023 as an observer model and Falcon-7B-Instruct as a performer model. Following (He et al., 2024; Spiegel and Macko, 2024; La Cava and Tagarelli, 2025), we train a Logistic Regressor on top of the extracted features to perform multiclass classification for the AA task.

Ensemble statistical approaches. To provide a stronger statistical approach, we combine nine statistical features into a statistical ensemble dubbed *StatEnsemble*. These are the metrics of Binoculars (Hans et al., 2024), Fast-DetectGPT (Bao et al., 2023), perplexity, Rank (Gehrmann et al., 2019), log-rank, log-likelihood, Entropy (Lavergne et al., 2008), LLM-Deviation (Wu and Xiang, 2023), and DetectLLM-LRR (Su et al., 2023), calculated based on mGPT-13B outputs. To perform a multiclass classification for AA, we train a Multi-layer Perceptron (MLP) classifier (MLP performing the best out of the examined Logistic Regressor, MLP, and Random Forest) with hyperparameters optimized using 5-fold grid search cross-validation over 1,000 steps. The remainder is kept to the default values of the `scikit-learn` library we used.

5.2 LLM-based Supervised Approaches

Fine-tuned encoders. For this type of detector, we consider RoBERTa-large (Liu et al., 2019) as an English-only pre-trained language model, and XLM-RoBERTa-large (Conneau et al., 2020) as the multilingual counterpart. Both models were fine-tuned for the AA task following (Wang et al., 2024a; Sarvazyan et al., 2023), with a learning rate of $2e$-6 and max sequence length of 512 tokens.

Contrastive learner. As a representative of contrastive approaches, we adapt the OTBDetector (La Cava and Tagarelli, 2025), which serves as the best-performing method in the recent *OpenTuringBench* benchmark for MGT attribution, to the multilingual AA task. It uses contrastive learning for fine-tuning a pre-trained model to separate latent representations of texts from different generators. For the multilingual setting, we replaced the original Longformer model with XLM-RoBERTa-large to ensure multilingual generalizability.

Fine-tuned decoder. We adapt the *mdok* detector (Macko, 2025), originally conceived as a multilingual binary MGT detection method, to the multilingual AA task. It is based on a fine-tuning of Qwen3-4B-Base (Team, 2025) model via QLoRA for enhancing generalization to out-of-distribution and obfuscated data, with a multiclass classification head performing multilingual classification. In addition, we have included Qwen3-4B-Base itself, fine-tuned similarly as the encoder models above.

[1] mGPT is a multilingual model with a similar architecture to the default GPT-J/Neo, outperforming other tested models in our experiments, including XLM-R, Qwen3-4B and 14B.

6 Experimental Setup

To address our research questions, we design four tasks that evaluate the feasibility and generalizability of multilingual authorship attribution. The first task corresponds to solving the ML-MGT problem (**RQ1**). To address the CL-MGT problem (**RQ2**), we distinguish between *per-language* and *per-language-family* cross-lingual transferability. The latter task corresponds to investigate the impact of the various LLM generators on the ML-MGT and CL-MGT performance (**RQ3**).

RQ1. Suitability of Existing Approaches to ML-MGT. To address RQ1, we evaluate the ability of the selected methods to handle the ML-MGT problem, by training them on data from all languages jointly, covering all 8 classes (7 LLM generators and human-authorship). The multilingual test set comprises the same languages, with performance reported as the macro-averaged F_1 score across all classes to ensure balanced treatment of each class regardless of frequency. Details on the train/test splits are shown in Table 1.

RQ2. Cross-lingual transferability of ML-MGT Authorship Attribution. We investigate whether AA methods trained on a single language or a combination of multiple languages could generalize their capabilities to other languages.

First, following (Macko et al., 2023, 2024, 2025a), we train AA methods on the subsets of English-, Spanish-, and Russian-only data from Table 1, using all 8 classes. Additionally, we train AA methods on a combination of English, Spanish, and Russian train data, which are sampled to $1/3$ each to ensure that these methods are trained on the same number of training samples as the monolingually trained AA methods. It should be noted that our choice of English, Spanish and Russian is motivated since they are the most popular languages with the two most representative scripts in MULTITUDE, i.e., Latin and Cyrillic.

To assess the *per-language transferability*, we evaluate the macro-averaged F_1 of AA methods on all the languages (including English, Spanish, and Russian), thus examining how a single language or language-subset during training can steer detectors to perform well in other languages, and comparing them to the multilingually trained detectors.

Similarly, to assess the *per-language-family transferability*, we investigate how the methods trained on one writing script can generalize to languages using a different script, hence to understand whether a language family plays a role in cross-lingual generalization. To this purpose, again we use English and Spanish to represent Latin-script training and Russian to represent Cyrillic-script training, and perform evaluation across all languages, measuring macro-averaged F_1.

By comparing intra-family and inter-family transfer performance, we aim to quantify whether family similarity/divergence affects the transferability of current AA methods in multilingual settings.

RQ3. Impact of LLM generators on the ML-MGT and CL-MGTperformance. We explore how the LLM generators influence the ML-MGT performance and cross-lingual generalization of attribution methods. To this aim, we examine language variations in class-level F_1 scores for each generator, shedding light on the interplay between generator identity and linguistic context in shaping adaptability and transferability.

7 Results

We present our experimental results in Sect. 7.1 for the RQ1 task, in Sect. 7.2 for the RQ2 tasks, and in Sect. 7.3 for the RQ3 task.

7.1 Multilingual Suitability Evaluation

Table 2 shows performance results (macro-averaged F_1) achieved by the methods in our ML-MGT problem setting based on 18 different languages. For reference, a random classifier performance has 0.125 macro F_1, due to distinguishing among 8 fully balanced classes.

At a first glance, we notice that five out of the eight detectors achieve macro $F_1 \geq 0.75$. Fine-tuning and contrastive approaches appear to help a lot in adaptability to the multilingual task, with the three best detectors, Qwen3-4B-Base, mdok and OTBDetector, remarkably showing an F_1 score above 0.9 in most cases, across all tested languages. Interestingly, OTBDetector appears to boost generalizability relatively better than mdok and Qwen3-4B-Base if we consider that, despite being $7\times$ smaller than them in parameter size, the F_1 score of OTBDetector only reduces by 3%, which might be due to a sharper decision boundary as determined by the contrastive loss used in OTBDetector.

As expected, detectors based on a multilingual pretraining (i.e., Qwen3-4B-Base and mdok, OTBDetector, XLM-RoBERTa-large) exhibit stronger

Lang. family →	Germanic			Romance			Slavic-Latin					Slavic-Cyrillic			Others				
Method ↓	de	en	nl	es	pt	ro	cs	hr	pl	sk	sl	bg	ru	uk	hu	el	ar	zh	all
Qwen3-4B-Base	**0.92**	0.91	**0.95**	**0.92**	**0.93**	**0.95**	**0.96**	0.95	0.94	0.97	0.95	**0.95**	**0.92**	0.93	**0.93**	0.93	**0.96**	0.85	**0.93**
mdok	0.92	**0.91**	0.95	0.91	0.93	0.94	0.95	**0.96**	**0.94**	**0.97**	**0.95**	0.93	0.91	**0.93**	0.93	**0.94**	0.96	**0.87**	0.93
OTBDetector	0.87	0.78	0.91	0.85	0.89	0.93	0.93	0.93	0.92	0.96	0.94	0.93	0.87	0.91	0.91	0.92	0.95	0.80	0.90
XLM-R-large	0.81	0.65	0.84	0.76	0.80	0.87	0.88	0.88	0.88	0.93	0.90	0.87	0.78	0.84	0.86	0.88	0.90	0.72	0.84
RoBERTa-large	0.78	0.72	0.81	0.74	0.80	0.84	0.83	0.83	0.81	0.85	0.84	0.63	0.63	0.67	0.76	0.59	0.70	0.60	0.75
StatEnsemble	0.49	0.33	0.55	0.45	0.47	0.48	0.43	0.43	0.50	0.43	0.31	0.51	0.48	0.48	0.50	0.41	0.40	0.35	0.45
Fast-DetectGPT	0.25	0.12	0.25	0.18	0.20	0.19	0.23	0.22	0.26	0.18	0.20	0.31	0.31	0.31	0.30	0.16	0.17	0.16	0.23
Binoculars	0.20	0.15	0.22	0.15	0.18	0.24	0.14	0.14	0.23	0.13	0.17	0.07	0.13	0.08	0.13	0.14	0.12	0.14	0.16
Average	0.65	0.57	0.68	0.62	0.65	0.68	0.67	0.67	0.69	0.68	0.66	0.65	0.63	0.64	0.66	0.62	0.64	0.56	0.65
Writing script →	Lat	Lat	Lat	Lat	Lat	Lat	Lat	Lat	Lat	Lat	Lat	Cyr	Cyr	Cyr	Lat	Grk	Arab	Han	

Table 2: **(RQ1)** Per-language macro-averaged F_1 scores of the selected methods on test data. Abbreviations of writing scripts are as follows: Lat = Latin, Cyr = Cyrillic, Grk = Greek, Arab = Arabic, Han = Hanzi. Bolded values indicate the best method for each test language. Darker shades of green indicate higher scores.

multilingual generalization compared to monolingual ones; however, it happens that English texts are generally difficult to attribute, even for English-only-pretrained methods like RoBERTa.

We ascribe this behavior to the fact that, since English is typically the best-supported language for most LLMs, the generator outputs might be harder to distinguish because they are more fluent and human-like, yet generators may converge stylistically, and differences between generators become subtle and blurry. In addition, XLM-RoBERTa-large performs worse than RoBERTa-large on the English portion of the multilingual test set, which can be explained since XLM-RoBERTa-large was originally pretrained on tens of languages simultaneously, and hence its capacity is spread across multiple languages, meaning its English representation is less specialized.

Finally, statistical approaches seem to be struggling overall across the results in Table 2. This is explained since they are conceived to simply separate human-written from machine-generated texts based on statistical patterns, which may not generalize to attribution. Furthermore, as most of these approaches rely on distributional patterns, their performance collapses in languages where LLM generators are very proficient—and thus adhere to human-like distributions—or in non-Latin scripts, which present distributional mismatches to Latin ones. Our conjecture is supported by the Binoculars case: it leverages the Falcon 7B model, which was trained mostly on English, German, Spanish, and French, i.e., Latin-script languages. Consequently, its representations are poorly suited to Cyrillic- or Arabic-script inputs, leading to failure in attributing texts in these languages.

7.2 Cross-lingual Transferability Evaluation

Language-level performance. Table 3 reveals several key findings at a language-level, which are summarized as follows. (We hereinafter leave Fast-DetectGPT and Binoculars out of evaluation given their poor performance as shown in Table 2).

Training on Russian (alone or in combination with others) has a significantly greater impact than other languages on the cross-lingual transferability, with +0.25 vs. English and +0.12 vs. Spanish in terms of overall best results; moreover, the observed benefit from training on Russian extends also to languages of a different family, especially non-Latin languages. By contrast, English appears to be the least generalizable, even among intra-script languages. This may be due to the simplicity of English tokenization and morphological structure, which fails to capture well to languages with richer morphological or syntactic complexity.

Focusing on the performance of the three best methods (i.e., Qwen3-4B-Base, mdok and OTB-Detector) on the results corresponding to English, Spanish, and Russian, respectively, a multilingual model from Table 2 appears to be preferable to a monolingual model trained on language L if the goal is to maximize the prediction performance on L only. At first glance, this might be seen as counterintuitive, since the inclusion of multiple languages in the training set could be expected to dilute language-specific patterns for L. However, the exposure to diverse linguistic structures may in fact enhance the model's generalization ability, even on individual languages. Nonetheless, the above remarks should be taken with a grain of salt, as differences in the number of training samples per language may introduce bias into the comparison.

	Lang. family →	Germanic			Romance			Slavic-Latin					Slavic-Cyrillic			Others				
	Method ↓	de	en	nl	es	pt	ro	cs	hr	pl	sk	sl	bg	ru	uk	hu	el	ar	zh	all
en	Qwen3-4B-Base	0.33	0.89	0.32	0.44	0.52	0.27	0.26	0.25	0.22	0.22	0.17	0.20	0.29	0.20	0.12	0.09	0.21	0.09	0.30
	mdok	0.50	**0.90**	0.39	**0.55**	**0.59**	0.34	0.31	0.29	0.32	0.20	0.22	0.13	0.22	0.18	0.26	0.10	0.10	0.12	0.36
	OTBDetector	0.51	0.83	0.42	0.47	0.53	**0.46**	**0.42**	**0.40**	**0.42**	**0.35**	**0.36**	**0.35**	**0.39**	**0.36**	**0.34**	**0.29**	**0.27**	**0.25**	**0.43**
	XLM-R-large	**0.52**	0.58	**0.43**	0.36	0.43	0.37	0.33	0.33	0.37	0.27	0.29	0.30	0.36	0.31	0.27	0.21	0.21	0.16	0.37
	RoBERTa-large	0.10	0.66	0.05	0.11	0.09	0.08	0.10	0.05	0.10	0.08	0.04	0.05	0.05	0.04	0.06	0.05	0.05	0.05	0.13
	StatEnsemble	0.21	0.53	0.20	0.27	0.23	0.10	0.09	0.11	0.10	0.07	0.09	0.08	0.13	0.02	0.11	0.03	0.16	0.19	0.16
es	Qwen3-4B-Base	**0.74**	**0.69**	**0.66**	**0.90**	**0.87**	0.66	**0.57**	0.44	0.66	**0.50**	0.43	0.41	0.57	0.47	0.43	0.23	0.32	0.12	**0.56**
	mdok	0.68	0.66	0.60	0.89	0.84	0.65	0.49	0.47	0.62	0.39	0.43	0.28	0.46	0.41	**0.43**	0.19	0.21	0.20	0.52
	OTBDetector	0.65	0.60	0.64	0.78	0.80	**0.69**	0.52	**0.50**	**0.67**	0.44	**0.44**	**0.48**	**0.57**	**0.52**	0.38	**0.36**	**0.40**	**0.32**	0.56
	XLM-R-large	0.57	0.39	0.54	0.58	0.55	0.40	0.39	0.32	0.39	0.33	0.32	0.36	0.37	0.34	0.31	0.26	0.25	0.21	0.41
	RoBERTa-large	0.48	0.20	0.53	0.58	0.60	0.60	0.35	0.45	0.25	0.24	0.22	0.05	0.04	0.04	0.19	0.06	0.07	0.07	0.32
	StatEnsemble	0.32	0.36	0.36	0.49	0.45	0.27	0.24	0.20	0.25	0.16	0.11	0.29	0.29	0.12	0.28	0.24	0.26	0.20	0.28
ru	Qwen3-4B-Base	0.64	0.40	**0.65**	0.61	0.64	0.72	0.77	0.73	0.74	**0.79**	0.76	0.80	0.87	0.77	0.50	0.44	**0.51**	0.26	0.67
	mdok	**0.73**	**0.49**	0.65	**0.63**	**0.72**	**0.72**	**0.80**	**0.75**	**0.77**	0.74	**0.79**	**0.80**	**0.88**	0.80	**0.70**	0.38	0.42	0.32	**0.68**
	OTBDetector	0.63	0.42	0.53	0.43	0.47	0.55	0.80	0.71	0.73	0.73	0.72	0.78	0.80	**0.83**	0.65	**0.49**	0.46	**0.45**	0.64
	XLM-R-large	0.43	0.23	0.30	0.30	0.30	0.44	0.73	0.59	0.62	0.65	0.67	0.67	0.63	0.69	0.64	0.40	0.43	0.37	0.53
	RoBERTa-large	0.07	0.05	0.06	0.07	0.07	0.08	0.06	0.09	0.07	0.09	0.08	0.38	0.43	0.38	0.06	0.22	0.15	0.09	0.17
	StatEnsemble	0.29	0.13	0.30	0.23	0.26	0.27	0.26	0.26	0.41	0.23	0.16	0.50	0.50	0.29	0.42	0.29	0.26	0.24	0.32
en-es-ru	Qwen3-4B-Base	**0.81**	0.84	0.79	0.81	0.84	**0.84**	**0.80**	0.73	0.77	**0.74**	**0.73**	**0.77**	0.80	0.72	**0.65**	**0.42**	0.43	0.31	0.72
	mdok	0.77	**0.88**	**0.79**	**0.86**	**0.86**	0.78	0.75	**0.73**	**0.80**	0.67	0.72	0.76	**0.85**	**0.78**	0.64	0.39	**0.46**	**0.35**	**0.72**
	OTBDetector	0.69	0.70	0.69	0.74	0.74	0.70	0.69	0.64	0.71	0.56	0.59	0.66	0.76	0.76	0.51	0.42	0.42	0.34	0.64
	XLM-R-large	0.57	0.44	0.47	0.48	0.46	0.53	0.64	0.59	0.61	0.56	0.63	0.59	0.56	0.56	0.60	0.36	0.41	0.34	0.53
	RoBERTa-large	0.46	0.57	0.54	0.48	0.56	0.55	0.47	0.49	0.32	0.29	0.22	0.37	0.40	0.39	0.23	0.25	0.18	0.15	0.40
	StatEnsemble	0.35	0.41	0.42	0.43	0.42	0.30	0.30	0.28	0.37	0.23	0.18	0.40	0.42	0.21	0.37	0.27	0.28	0.22	0.34
	Writing script →	Lat	Lat	Lat	Lat	Lat	Lat	Lat	Lat	Lat	Lat	Lat	Cyr	Cyr	Cyr	Lat	Grk	Arab	Han	

Table 3: **(RQ2)** Per-language cross-lingual macro-averaged F_1 scores of the selected methods on test data. Writing scripts are as follows: Lat = Latin, Cyr = Cyrillic, Grk = Greek, Arab = Arabic, Han = Hanzi. Bolded values indicate the best method for each training-language and test-language pair. Darker shades of green indicate higher scores.

Language-family-level performance. Aggregating results from Table 3 by language family (cf. Table 6 in the Appendix D) reveals the beneficial effect of Russian on cross-lingual transferability. Notably, training on Russian alone yields optimal performance in six out of eight test families, i.e., all except Germanic and Romance. For these two families, combining Russian with English and Spanish is essential to maximize performance.

Why Russian languages support better cross-lingual transferability. We attribute the stronger cross-lingual transferability observed with Russian to a number of syntactic and morphological properties of the language (Dryer and Haspelmath, 2013). Russian is rich in morphology, with a high inflectional structure, where grammatical roles (e.g., subject, object, verb) are encoded via an extensive use of word endings that allow words to convey a wide range of meanings within a sentence (Iggesen, 2013; Bickel and Nichols, 2013). This contrasts with English, where discourse construction typically relies on fixed syntax and a simpler morphology. Consequently, its morphological richness may encourage models trained on Russian to capture deeper linguistic signals that transfer more robustly across languages, whereas models trained on English might learn more superficial token-level rules that do not generalize well. This observation is further supported by experiments in Table 7, referring the results of the selected detection methods fine-tuned using the other Slavic languages. They show a similar superiority as the Russian language.

7.3 Influence of LLM Generators on ML-MGT and CL-MGT

We analyze the influence that the various LLM generators have on multilingual authorship attribution and its cross-lingual transferability by examining the error patterns of mdok and OTBDetector, which here are selected as they have shown the best trade-off between efficiency and generalizability.

ML-MGT Patterns. Both mdok and OTBDetector exhibit very high attribution-performance when trained and evaluated on the full set of available languages, confirming the remarkable results from Table 2. Their confusion matrices suggest that those detectors can effectively learn the stylistic footprint of each LLM generator and generalize attribution across languages. A closer look at the per-generator behavior (cf. Table 8 in Appendix E)

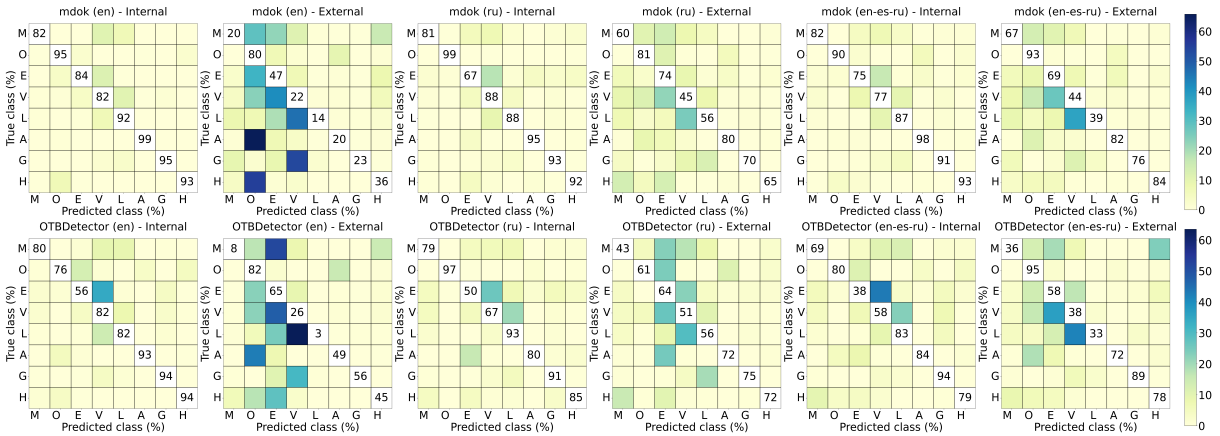

Figure 2: **(RQ3)** Confusion matrices (row-wise percentages) for two of the best-performing approaches, i.e., mdok (top) and OTBDetector (bottom), by varying the training data. *Internal* and *External* here indicate that the method has been evaluated on the same language as training and on all but the training language, respectively. Numbers in the diagonal indicate the percentage of correct predictions. LLM generators are referred to using the first letter, namely M = Mistral, O = OPT, E = Eagle, V = Vicuna, L = Llama2, A = Aya, G = GPT-3.5, and H = human.

reveals that the detectors struggle slightly more in attributing models like Mistral, Eagle, and Vicuna, while excelling in the Aya, GPT-3.5, human, and OPT classes. These differences are also language-dependent, with Cyrillic languages, Hungarian, Chinese, Czech, and even English showing higher error rates.

CL-MGT Patterns. Figure 2 provides the confusion matrices of the two of the best detectors based on the language selection for our RQ2 (i.e., en, ru, en-es-ru). Here we distinguish between an *Internal* setting, where training and test languages are the same, and an *External* setting, where the test languages are missing in the training set. In the former case, both mdok and OTBDetector perform well across the three language-group scenarios; however, under the External setting, performance tends to worsen, with increasing confusion among architecturally similar models.

Error Trends by Generator. Llama2-70B and Vicuna-13B appear to be relatively difficult to attribute, especially in the English-based External setting, which might be due to the shared underlying architecture among these models—Vicuna is in fact a further-fine-tuning of Llama. Interestingly, human-written texts are among the easiest to attribute, suggesting that despite the fluency LLMs have in producing multilingual texts, distinct human-specific patterns remain detectable. Finally, OPT-30B and Eagle-7B emerge as the "catch-all" classes for English and Russian, respectively, in the External setting, as a recurring pattern for both mdok and OTBDetector involves overpredicting those LLMs. We ascribe this to the tendency of the two LLMs to generate texts with fewer stylistic variations, thus becoming the most predictable classes when the detector is uncertain—especially under the CL-MGT problem.

8 Conclusions

Despite the growing multilingual usage of LLMs, current efforts in authorship attribution of MGT remain largely confined to monolingual contexts, particularly English. In this work, we filled this gap by formally defining and exploring the problems of multilingual and cross-lingual authorship attribution in MGT. We evaluated the performance of established authorship attribution approaches in the multilingual setting, as well as their ability to generalize across languages. Our experiments, covering 18 languages and 8 author classes (7 LLMs and a human class), demonstrate that while some existing methods can be adapted to multilingual authorship attribution, their effectiveness varies widely, highlighting the challenge of cross-lingual transferability and the need for further development in the field for real-world multilingual usage.

Code and data resources associated with this work are available at https://github.com/MLNTeam-Unical/Multilingual-MGT-AA.

Future work. Although the chosen news-domain offers a significant testbed due to its linguistic and topical diversity, our investigation scope should be extended. While we have already obtained preliminary results regarding social contents, as discussed in Appendix C, further analysis on other domains is

left for future work. Additionally, we would like to expand our analysis and findings geographically by considering more (low-resource) languages. Also, we aim to investigate the impact of adversarial attacks on the accuracy and cross-lingual transferability of multilingual MGT attribution approaches.

Limitations

Language Coverage. Our study considers 18 languages from the MULTITUDE dataset. While being representative of a broad range of language families and scripts, it does not fully represent the linguistic diversity encountered in real-world settings. For this reason, we are committed to investigating multilingual authorship attribution in more (low-resource) languages.

Domain Coverage. Our study deliberately focuses on the news domain, as it represents one of the most challenging settings due to high variability in topics, styles, and linguistic registers across languages. While these aspects make attribution intrinsically harder and provide a rigorous testbed for real-world applicability, we acknowledge that attribution performance may differ in other domains (cf. Appendix C for insights on social content) and our findings may not strongly generalize beyond news-style content. Expanding our investigations to a broader set of domains is part of future work.

LLM-generators Coverage. Our study considers a fixed and controlled set of seven LLM generators and one human-authored class. While these models cover a broad range of architectures and sizes, we acknowledge that the rapidly evolving landscape of NLP might introduce advancements that may degrade attribution methods when faced with unseen or future models, e.g., different data distributions or new decoding strategies.

Adversarial Attacks. Our study assumes clean, well-formed text and does not account for adversarial manipulations (e.g., paraphrasing, prompting, obfuscation) aimed at challenging detection and attribution systems. Addressing these scenarios remains an important direction for future work.

Ethical Considerations

Broader Impact. Authorship Attribution systems can be misused or overtrusted, with unintended effects on society. For this reason, we urge all stakeholders to use such methodologies responsibly, keeping in account potential biases across languages (e.g., higher false positives for low-resource languages/scripts), and ensuring humans-in-the-loop while using such tools for decision making.

Acknowledgments

This work was partially supported by the European Union under the Horizon Europe project AI-CODE, GA No. 101135437; by the EU NextGenerationEU through the Recovery and Resilience Plan for Slovakia under the project No. 09I01-03-V04-00059; by the project "Future Artificial Intelligence Research (FAIR)" spoke 9 (H23C22000860006), and the project SERICS (PE00000014).

We acknowledge the EuroHPC Joint Undertaking for awarding us access to Leonardo at CINECA, Italy.

References

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, and 1 others. 2023. The falcon series of open language models. *arXiv preprint arXiv:2311.16867*.

Guangsheng Bao, Yanbin Zhao, Zhiyang Teng, Linyi Yang, and Yue Zhang. 2023. Fast-DetectGPT: Efficient zero-shot detection of machine-generated text via conditional probability curvature. In *The Twelfth International Conference on Learning Representations*.

Amrita Bhattacharjee, Tharindu Kumarage, Raha Moraffah, and Huan Liu. 2023. ConDA: Contrastive domain adaptation for AI-generated text detection. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 598–610, Nusa Dua, Bali. Association for Computational Linguistics.

Amrita Bhattacharjee and Huan Liu. 2024. Fighting Fire with Fire: Can ChatGPT Detect AI-generated Text? *SIGKDD Explor. Newsl.*, 25(2):14–21.

Amrita Bhattacharjee, Raha Moraffah, Joshua Garland, and Huan Liu. 2024. Eagle: A domain generalization framework for ai-generated text detection. *arXiv preprint arXiv:2403.15690*.

Balthasar Bickel and Johanna Nichols. 2013. Inflectional synthesis of the verb (v2020.4). In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Zenodo.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg,

and 1 others. 2023. Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv preprint arXiv:2303.12712*.

Canyu Chen and Kai Shu. 2024. Can llm-generated misinformation be detected? In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Evan Crothers, Nathalie Japkowicz, and Herna L Viktor. 2023. Machine-generated text: A comprehensive survey of threat models and detection methods. *IEEE Access*.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online (v2020.4)*. Zenodo.

Liam Dugan, Alyssa Hwang, Filip Trhlík, Andrew Zhu, Josh Magnus Ludan, Hainiu Xu, Daphne Ippolito, and Chris Callison-Burch. 2024. 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)*, pages 12463–12492, Bangkok, Thailand. Association for Computational Linguistics.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander Rush. 2019. GLTR: Statistical detection and visualization of generated text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Association for Computational Linguistics.

Xun Guo, Yongxin He, Shan Zhang, Ting Zhang, Wanquan Feng, Haibin Huang, and Chongyang Ma. 2024. Detective: Detecting AI-generated text via multi-level contrastive learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Abhimanyu Hans, Avi Schwarzschild, Valeriia Cherepanova, Hamid Kazemi, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2024. Spotting llms with binoculars: Zero-shot detection of machine-generated text. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*.

Xinlei He, Xinyue Shen, Zeyuan Chen, Michael Backes, and Yang Zhang. 2024. Mgtbench: Benchmarking machine-generated text detection. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, CCS '24, page 2251–2265, New York, NY, USA. Association for Computing Machinery.

Baixiang Huang, Canyu Chen, and Kai Shu. 2025. Authorship attribution in the era of llms: Problems, methodologies, and challenges. *SIGKDD Explor. Newsl.*, 26(2):21–43.

Oliver A. Iggesen. 2013. Number of cases (v2020.4). In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Zenodo.

Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, and Douglas Eck. 2020. Automatic detection of generated text is easiest when humans are fooled. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1808–1822, Online. Association for Computational Linguistics.

Maurice Jakesch, Jeffrey T. Hancock, and Mor Naaman. 2023. Human heuristics for ai-generated language are flawed. *Proceedings of the National Academy of Sciences*, 120(11):e2208839120.

Ganesh Jawahar, Muhammad Abdul-Mageed, and Laks Lakshmanan, V.S. 2020. Automatic detection of machine generated text: A critical survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2296–2309, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Zae Myung Kim, Kwang Lee, Preston Zhu, Vipul Raheja, and Dongyeop Kang. 2024. Threads of subtlety: Detecting machine-generated texts through discourse motifs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5449–5474, Bangkok, Thailand. Association for Computational Linguistics.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. 2023. A watermark for large language models. In *Proc. of Int. Conf. on Machine Learning (ICML)*, pages 17061–17084.

Lucio La Cava, Davide Costa, and Andrea Tagarelli. 2024. Is contrasting all you need? contrastive learning for the detection and attribution of ai-generated text. In *ECAI 2024 - 27th European Conference on Artificial Intelligence, 19-24 October 2024, Santiago de Compostela, Spain*, volume 392 of *Frontiers in Artificial Intelligence and Applications*, pages 3179–3186. IOS Press.

Lucio La Cava and Andrea Tagarelli. 2025. OpenTuringBench: An open-model-based benchmark and framework for machine-generated text detection and attribution. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 26666–26682, Suzhou, China. Association for Computational Linguistics.

Thomas Lavergne, Tanguy Urvoy, and François Yvon. 2008. Detecting fake content with relative entropy scoring. In *Proceedings of the 2008 International Conference on Uncovering Plagiarism, Authorship*

and Social Software Misuse - Volume 377, PAN'08, page 27–31, Aachen, DEU. CEUR-WS.org.

Xiaoze Liu, Ting Sun, Tianyang Xu, Feijie Wu, Cunxiang Wang, Xiaoqian Wang, and Jing Gao. 2024. SHIELD: Evaluation and defense strategies for copyright compliance in LLM text generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1640–1670, Miami, Florida, USA. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Dominik Macko. 2025. mdok of KInIT: Robustly fine-tuned llm for binary and multiclass ai-generated text detection. *arXiv preprint arXiv:2506.01702*.

Dominik Macko, Jakub Kopál, Robert Moro, and Ivan Srba. 2025a. 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)*, pages 727–752, Vienna, Austria. Association for Computational Linguistics.

Dominik Macko, Jakub Kopal, Robert Moro, and Ivan Srba. 2025b. Multitudev3. *Zenodo*, 15519413.

Dominik Macko, Robert Moro, Adaku Uchendu, Jason Lucas, Michiharu Yamashita, Matúš Pikuliak, Ivan Srba, Thai Le, Dongwon Lee, Jakub Simko, and Maria Bielikova. 2023. MULTITuDE: Large-scale multilingual machine-generated text detection benchmark. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9960–9987, Singapore. Association for Computational Linguistics.

Dominik Macko, Robert Moro, Adaku Uchendu, Ivan Srba, Jason S Lucas, Michiharu Yamashita, Nafis Irtiza Tripto, Dongwon Lee, Jakub Simko, and Maria Bielikova. 2024. Authorship obfuscation in multilingual machine-generated text detection. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6348–6368, Miami, Florida, USA. Association for Computational Linguistics.

Yibo Miao, Hongcheng Gao, Hao Zhang, and Zhijie Deng. 2024. Efficient detection of LLM-generated texts with a Bayesian surrogate model. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6118–6130, Bangkok, Thailand. Association for Computational Linguistics.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. 2023. Detectgpt: Zero-shot machine-generated text detection using probability curvature. In *Proc. of Int. Conf. on Machine Learning (ICML)*, pages 24950–24962. PMLR.

Areg Mikael Sarvazyan, José Ángel González, Marc Franco-Salvador, Francisco Rangel, Berta Chulvi, and Paolo Rosso. 2023. Overview of autextification at iberlef 2023: Detection and attribution of machine-generated text in multiple domains. *Procesamiento del Lenguaje Natural*, 71:275–288.

Tatiana Shamardina, Vladislav Mikhailov, Daniil Chernianskii, Alena Fenogenova, Marat Saidov, Anastasiya Valeeva, Tatiana Shavrina, Ivan Smurov, Elena Tutubalina, and Ekaterina Artemova. 2022. Findings of the the RuATD Shared Task 2022 on artificial text detection in Russian. In *Computational Linguistics and Intellectual Technologies*. RSUH.

Oleh Shliazhko, Alena Fenogenova, Maria Tikhonova, Anastasia Kozlova, Vladislav Mikhailov, and Tatiana Shavrina. 2024. mGPT: Few-shot learners go multilingual. *Transactions of the Association for Computational Linguistics*, 12:58–79.

Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, and 1 others. 2019. Release strategies and the social impacts of language models. *arXiv preprint arXiv:1908.09203*.

Michal Spiegel and Dominik Macko. 2024. IMGTB: A framework for machine-generated text detection benchmarking. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 172–179, Bangkok, Thailand. Association for Computational Linguistics.

Jinyan Su, Terry Zhuo, Di Wang, and Preslav Nakov. 2023. DetectLLM: Leveraging log rank information for zero-shot detection of machine-generated text. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12395–12412, Singapore. Association for Computational Linguistics.

Ruixiang Tang, Yu-Neng Chuang, and Xia Hu. 2024. The science of detecting llm-generated text. *Communications of the ACM*, 67(4):50–59.

Zhen Tao, Yanfang Chen, Dinghao Xi, Zhiyu Li, and Wei Xu. 2024. Towards reliable detection of LLM-generated texts: A comprehensive evaluation framework with CUDRT. *arXiv preprint arXiv:2406.09056*.

Qwen Team. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Eduard Tulchinskii, Kristian Kuznetsov, Kushnareva Laida, Daniil Cherniavskii, Sergey Nikolenko, Evgeny Burnaev, Serguei Barannikov, and Irina Piontkovskaya. 2023. Intrinsic dimension estimation for robust detection of AI-generated texts. In *Proc. of Conf. on Advances in Neural Information Processing Systems (NIPS)*.

Adaku Uchendu, Thai Le, and Dongwon Lee. 2023a. Attribution and obfuscation of neural text authorship:

A data mining perspective. *SIGKDD Explor. Newsl.*, 25(1):1–18.

Adaku Uchendu, Thai Le, and Dongwon Lee. 2023b. Toproberta: Topology-aware authorship attribution of deepfake texts. *arXiv preprint arXiv:2309.12934*.

Adaku Uchendu, Thai Le, Kai Shu, and Dongwon Lee. 2020. Authorship attribution for neural text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8384–8395, Online. Association for Computational Linguistics.

Daniel Varab and Natalie Schluter. 2021. MassiveSumm: a very large-scale, very multilingual, news summarisation dataset. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10150–10161, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Christoforos Vasilatos, Manaar Alam, Talal Rahwan, Yasir Zaki, and Michail Maniatakos. 2023. HowkGPT: Investigating the detection of ChatGPT-generated university student homework through context-aware perplexity analysis. *arXiv preprint arXiv:2305.18226*.

Saranya Venkatraman, Adaku Uchendu, and Dongwon Lee. 2024. GPT-who: An information density-based machine-generated text detector. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 103–115, Mexico City, Mexico. Association for Computational Linguistics.

Vivek Verma, Eve Fleisig, Nicholas Tomlin, and Dan Klein. 2024. Ghostbuster: Detecting text ghostwritten by large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1702–1717, Mexico City, Mexico. Association for Computational Linguistics.

Pengyu Wang, Linyang Li, Ke Ren, Botian Jiang, Dong Zhang, and Xipeng Qiu. 2023. SeqXGPT: Sentence-level AI-generated text detection. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1144–1156, Singapore. Association for Computational Linguistics.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Osama Mohammed Afzal, Tarek Mahmoud, Giovanni Puccetti, and Thomas Arnold. 2024a. SemEval-2024 task 8: Multidomain, multimodel and multilingual machine-generated text detection. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 2057–2079, Mexico City, Mexico. Association for Computational Linguistics.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Osama Mohammed Afzal, Tarek Mahmoud, Giovanni Puccetti, Thomas Arnold, Alham Aji, Nizar Habash, Iryna Gurevych, and Preslav Nakov. 2024b. M4GT-bench: Evaluation benchmark for black-box machine-generated text detection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3964–3992, Bangkok, Thailand. Association for Computational Linguistics.

Junchao Wu, Shu Yang, Runzhe Zhan, Yulin Yuan, Lidia Sam Chao, and Derek Fai Wong. 2025. A survey on LLM-generated text detection: Necessity, methods, and future directions. *Computational Linguistics*, 51(1):275–338.

Zhendong Wu and Hui Xiang. 2023. MFD: Multi-feature detection of LLM-generated text. *PREPRINT (Version 1) available at Research Square*.

Xiaojun Xu, Yuanshun Yao, and Yang Liu. 2024. Learning to watermark llm-generated text via reinforcement learning. *arXiv preprint arXiv:2403.10553*.

KiYoon Yoo, Wonhyuk Ahn, Jiho Jang, and Nojun Kwak. 2023. Robust multi-bit natural language watermarking through invariant features. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2092–2115, Toronto, Canada. Association for Computational Linguistics.

Aneta Zugecova, Dominik Macko, Ivan Srba, Robert Moro, Jakub Kopál, Katarína Marcinčinová, and Matúš Mesarčík. 2025. Evaluation of LLM vulnerabilities to being misused for personalized disinformation generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 780–797, Vienna, Austria. Association for Computational Linguistics.

A Computational Resources

For fine-tuning and inference of authorship attribution methods (a single run for each version of fine-tuned authorship attribution method), as well as for hyperparameters optimization, we used a machine allocated with 8 CPU cores (Intel Xeon Platinum 8358 CPU, 2.6 GHz), 128GB RAM, and 1× A100 64GB GPU, cumulatively consuming approximately 500 GPU-hours.

B Multilingual MGT Datasets

In Table 4, we summarize the basic statistics about generators, languages, and domains of datasets that can be regarded as potentially useful for multilingual authorship attribution.

CUDRT and RAID-extra cover 3 or fewer languages, thus being not well-suited to our cross-lingual study. M4GT-Bench is a composition of multiple datasets covering various domains, which might introduce a bias in the results due to the different nature of the domains.

To the best of our knowledge, the MULTITUDE collection is the only one containing a relevant set of generators, text-generation settings, and domains for each language, enabling a proper cross-lingual transferability evaluation—especially in its latest version.

It should be noted that MultiSocial also offers broad coverage in terms of both languages and generators. However, it contains an uneven number of samples across languages, and the data are drawn from multiple social media platforms. This heterogeneity may introduce inconsistencies in writing style and topic distribution. These considerations motivated our choice of MULTITUDE as the target dataset for our study. Nonetheless, as discussed in Appendix C, we additionally leverage MultiSocial for a preliminary investigation under *out-of-domain conditions*, in order to assess the robustness and generalizability of authorship attribution methods beyond the in-domain (i.e., MULTITUDE) setting.

C Out-of-Domain Generalization

To evaluate the generalization of the authorship attribution methods reported in Table 2 to unseen domains, we evaluated them on the MultiSocial data. Indeed, not only it covers a different domain (i.e., social), but it also contains 4 additional languages besides the 18 languages we used for training. Results, which are shown in Table 5, highlight that OTBDetector and XLM-RoBERTa-large achieve better average macro F_1 than the other methods, although, as expected, overall performance remains substantially lower across all approaches w.r.t. the in-domain setting.

D Per-language-family cross-lingual performance

Table 6 provides details on the per-language-family cross-lingual transferability by aggregating the single-language cross-lingual transferability results from Table 3 in the main text by language family.

Table 7 provides cross-lingual transferability of three selected detection methods fine-tuned on individual Slavic languages, analogously to Table 3. We can see a similarly strong transferability as in case of the Russian language.

E Per-generator multi-lingual and cross-lingual performance

The finer-granularity multilingual (for each test language) results per-class (i.e., generator) of the selected authorship attribution methods are provided in Table 8. In this single-class evaluation scenario, the performance is reported in the form of a weighted average F_1 score (since non-evaluated classes have no supporting samples). Analogously, Table 9 reports per-generator performance of two of the best (mdok and OTBDetector) authorship attribution methods for cross-lingual experiments.

Dataset	Reference	#Generators	#Languages	#Domains
CUDRT	(Tao et al., 2024)	5	2	6
M4GT-Bench	(Wang et al., 2024b)	8	9	6
MULTITuDE_v1	(Macko et al., 2023)	9	11	1
MULTITuDE_v3	(Macko et al., 2025b)	8	21	1
MultiSocial	(Macko et al., 2025a)	8	22	1
RAID-extra	(Dugan et al., 2024)	11	3	8

Table 4: Overview of existing resources for multilingual machine-generated text detection.

Lang. family →	Germanic			Romance				Slavic-Latin					Slavic-Cyrillic			Others							
Method ↓	de	en	nl	ca*	es	pt	ro	cs	hr	pl	sk	sl	bg	ru	uk	ga*	gd*	et*	hu	el	ar	zh	all
OTBDetector	**0.36**	**0.32**	0.18	**0.25**	**0.21**	**0.29**	**0.19**	**0.21**	0.16	0.16	0.15	0.23	0.25	0.24	0.22	**0.23**	**0.23**	**0.28**	**0.33**	**0.35**	0.29	0.36	**0.24**
XLM-R-large	0.20	0.14	**0.21**	0.17	0.17	0.17	0.17	0.19	**0.18**	**0.18**	**0.28**	**0.28**	**0.28**	**0.26**	**0.27**	0.15	0.15	0.14	0.19	0.26	**0.34**	0.35	0.21
mdok	0.18	0.18	0.18	0.16	0.17	0.18	0.16	0.19	0.17	0.16	0.23	0.23	0.24	0.25	0.22	0.09	0.10	0.11	0.19	0.24	0.30	0.40	0.20
Qwen3-4B-Base	0.18	0.19	0.19	0.16	0.18	0.17	0.13	0.18	0.15	0.15	0.26	0.22	0.27	0.22	0.20	0.11	0.10	0.10	0.16	0.21	0.32	**0.41**	0.19
RoBERTa-large	0.14	0.15	0.14	0.12	0.13	0.13	0.13	0.15	0.17	0.12	0.17	0.20	0.07	0.09	0.06	0.09	0.12	0.10	0.12	0.14	0.12	0.28	0.14
StatEnsemble	0.14	0.13	0.14	0.13	0.13	0.13	0.11	0.12	0.11	0.13	0.13	0.09	0.16	0.15	0.13	0.07	0.06	0.06	0.12	0.16	0.17	0.13	0.13
Fast-DetectGPT	0.11	0.12	0.10	0.11	0.11	0.11	0.09	0.12	0.11	0.10	0.08	0.10	0.11	0.13	0.12	0.12	0.16	0.08	0.08	0.08	0.09	0.11	0.11
Binoculars	0.11	0.10	0.10	0.07	0.11	0.09	0.08	0.08	0.09	0.10	0.07	0.09	0.09	0.13	0.08	0.08	0.06	0.06	0.07	0.13	0.13	0.10	0.10
Average	0.18	0.17	0.15	0.15	0.15	0.16	0.13	0.16	0.14	0.14	0.17	0.18	0.18	0.18	0.16	0.12	0.12	0.12	0.16	0.19	0.22	0.27	0.16
Writing script →	Lat	Lat	Lat	Lat	Lat	Lat	Lat	Lat	Lat	Lat	Lat	Lat	Cyr	Cyr	Cyr	Lat	Lat	Lat	Lat	Grk	Arab	Han	

Table 5: Per-language macro-averaged F_1 scores of the selected methods on MultiSocial test data (i.e., generalization to social-media domain). Abbreviations of writing scripts are as follows: Lat = Latin, Cyr = Cyrillic, Grk = Greek, Arab = Arabic, Han = Hanzi. Bolded values indicate the best method for each test language. Darker shades of green indicate higher scores. * denotes the new languages.

	Lang. family →	Germanic	Romance	Slavic-Latin	Slavic-Cyrillic	Uralic	Greek	Semitic	Sino-Tibetan
	Method ↓	($N=3$)	($N=3$)	($N=5$)	($N=3$)	($N=1$)	($N=1$)	($N=1$)	($N=1$)
en	Qwen3-4B-Base	0.52	0.41	0.23	0.23	0.12	0.09	0.21	0.09
	mdok	0.60	0.49	0.27	0.18	0.26	0.10	0.10	0.12
	OTBDetector	0.58	0.49	0.39	0.36	0.34	0.29	0.27	0.25
	XLM-R-large	0.51	0.39	0.31	0.32	0.27	0.21	0.21	0.16
	RoBERTa-large	0.27	0.09	0.07	0.04	0.06	0.05	0.05	0.05
	StatEnsemble	0.31	0.20	0.09	0.08	0.11	0.03	0.16	0.19
es	Qwen3-4B-Base	0.70	0.81	0.52	0.48	0.43	0.23	0.32	0.12
	mdok	0.65	0.79	0.48	0.38	0.43	0.19	0.21	0.20
	OTBDetector	0.63	0.75	0.51	0.52	0.38	0.36	0.40	0.32
	XLM-R-large	0.50	0.51	0.35	0.36	0.31	0.26	0.25	0.21
	RoBERTa-large	0.40	0.59	0.30	0.04	0.19	0.06	0.07	0.07
	StatEnsemble	0.35	0.40	0.19	0.23	0.28	0.24	0.26	0.20
ru	Qwen3-4B-Base	0.56	0.65	0.76	0.81	0.50	0.44	**0.51**	0.26
	mdok	0.62	0.69	**0.77**	**0.83**	**0.70**	0.38	0.42	0.32
	OTBDetector	0.53	0.48	0.74	0.80	0.65	**0.49**	0.46	**0.45**
	XLM-R-large	0.32	0.35	0.65	0.66	0.64	0.40	0.43	0.37
	RoBERTa-large	0.06	0.07	0.08	0.40	0.06	0.22	0.15	0.09
	StatEnsemble	0.24	0.25	0.26	0.43	0.42	0.29	0.26	0.24
en-es-ru	Qwen3-4B-Base	**0.81**	0.83	0.76	0.76	0.65	0.42	0.43	0.31
	mdok	0.81	**0.84**	0.74	0.80	0.64	0.39	0.46	0.35
	OTBDetector	0.69	0.73	0.64	0.72	0.51	0.42	0.42	0.34
	XLM-R-large	0.49	0.49	0.61	0.57	0.60	0.36	0.41	0.34
	RoBERTa-large	0.52	0.53	0.36	0.39	0.23	0.25	0.18	0.15
	StatEnsemble	0.39	0.39	0.27	0.34	0.37	0.27	0.28	0.22

Table 6: **(RQ2)** Per-language-family cross-lingual performance (macro F_1) of the selected methods on test data. Rows are grouped by training language. N denotes the number of test languages belonging to the language family, from which the mean value is calculated. Bolded values correspond to the best results per test-language-group.

	Lang. family →	Germanic			Romance			Slavic-Latin					Slavic-Cyrillic			Others				
	Method ↓	de	en	nl	es	pt	ro	cs	hr	pl	sk	sl	bg	ru	uk	hu	el	ar	zh	all
ru	Qwen3-4B-Base	0.64	0.40	0.65	0.61	0.64	0.72	0.77	0.73	0.74	0.79	0.76	0.80	0.87	0.77	0.50	0.44	0.51	0.26	0.67
	mdok	0.73	0.49	0.65	0.63	0.72	0.72	0.80	0.75	0.77	0.74	0.79	0.80	0.88	0.80	0.70	0.38	0.42	0.32	0.68
	XLM-R-large	0.43	0.23	0.30	0.30	0.30	0.44	0.73	0.59	0.62	0.65	0.67	0.67	0.63	0.69	0.64	0.40	0.43	0.37	0.53
bg	Qwen3-4B-Base	0.53	0.49	0.58	0.56	0.59	0.73	0.77	0.77	0.73	0.77	0.77	0.91	0.72	0.81	0.67	0.53	0.54	0.24	0.67
	mdok	0.60	0.37	0.58	0.60	0.64	0.73	0.80	0.76	0.75	0.79	0.83	0.93	0.78	0.80	0.75	0.50	0.31	0.24	0.68
	XLM-R-large	0.40	0.24	0.39	0.34	0.37	0.50	0.69	0.62	0.61	0.71	0.75	0.74	0.53	0.63	0.68	0.47	0.51	0.37	0.55
uk	Qwen3-4B-Base	0.49	0.33	0.62	0.51	0.57	0.66	0.72	0.69	0.71	0.73	0.69	0.76	0.68	0.89	0.51	0.43	0.44	0.17	0.61
	mdok	0.64	0.47	0.69	0.66	0.70	0.77	0.82	0.79	0.80	0.73	0.75	0.83	0.83	0.92	0.70	0.48	0.36	0.24	0.69
	XLM-R-large	0.37	0.25	0.36	0.28	0.30	0.43	0.68	0.57	0.58	0.67	0.69	0.67	0.55	0.67	0.64	0.43	0.48	0.37	0.52
cs	Qwen3-4B-Base	0.65	0.37	0.61	0.52	0.63	0.79	0.92	0.80	0.78	0.90	0.83	0.66	0.58	0.64	0.67	0.37	0.48	0.18	0.66
	mdok	0.69	0.32	0.61	0.47	0.58	0.74	0.94	0.81	0.83	0.85	0.83	0.49	0.59	0.62	0.74	0.24	0.27	0.23	0.63
	XLM-R-large	0.48	0.23	0.41	0.33	0.35	0.48	0.76	0.59	0.63	0.77	0.73	0.64	0.55	0.61	0.66	0.36	0.42	0.29	0.54
hr	Qwen3-4B-Base	0.67	0.24	0.66	0.49	0.56	0.73	0.77	0.94	0.78	0.72	0.53	0.67	0.49	0.55	0.62	0.34	0.38	0.16	0.60
	mdok	0.71	0.42	0.64	0.54	0.59	0.79	0.76	0.95	0.73	0.71	0.32	0.63	0.58	0.52	0.65	0.31	0.24	0.19	0.59
	XLM-R-large	0.39	0.13	0.38	0.23	0.28	0.50	0.63	0.76	0.57	0.51	0.15	0.56	0.44	0.50	0.58	0.36	0.34	0.28	0.45
pl	Qwen3-4B-Base	0.64	0.49	0.67	0.56	0.63	0.68	0.73	0.72	0.89	0.71	0.69	0.57	0.50	0.57	0.55	0.38	0.43	0.18	0.61
	mdok	0.71	0.39	0.73	0.55	0.64	0.80	0.80	0.80	0.92	0.74	0.80	0.61	0.61	0.64	0.71	0.35	0.27	0.21	0.65
	XLM-R-large	0.51	0.26	0.48	0.39	0.42	0.57	0.68	0.63	0.77	0.72	0.68	0.64	0.51	0.59	0.68	0.44	0.45	0.41	0.56
sk	Qwen3-4B-Base	0.49	0.41	0.54	0.48	0.61	0.69	0.76	0.65	0.71	0.97	0.67	0.61	0.57	0.57	0.56	0.46	0.58	0.28	0.61
	mdok	0.57	0.31	0.54	0.42	0.53	0.64	0.76	0.71	0.72	0.96	0.74	0.57	0.51	0.54	0.67	0.35	0.32	0.23	0.59
	XLM-R-large	0.40	0.19	0.29	0.27	0.39	0.33	0.52	0.40	0.46	0.89	0.45	0.49	0.49	0.43	0.41	0.48	0.48	0.24	0.45
sl	Qwen3-4B-Base	0.54	0.35	0.60	0.52	0.64	0.70	0.83	0.80	0.72	0.82	0.95	0.71	0.54	0.60	0.70	0.38	0.36	0.16	0.63
	mdok	0.72	0.43	0.68	0.51	0.55	0.68	0.76	0.76	0.69	0.72	0.95	0.57	0.51	0.53	0.62	0.27	0.21	0.18	0.60
	XLM-R-large	0.39	0.25	0.37	0.33	0.34	0.49	0.62	0.67	0.55	0.70	0.76	0.63	0.45	0.50	0.61	0.39	0.39	0.29	0.51
	Writing script →	Lat	Lat	Lat	Lat	Lat	Lat	Lat	Lat	Lat	Lat	Lat	Cyr	Cyr	Cyr	Lat	Grk	Arab	Han	

Table 7: **(RQ2)** Per-language cross-lingual macro-averaged F_1 scores of the selected methods fine-tuned on Slavic languages on test data. Writing scripts are as follows: Lat = Latin, Cyr = Cyrillic, Grk = Greek, Arab = Arabic, Han = Hanzi. Bolded values indicate the best method for each training-language and test-language pair. Darker shades of green indicate higher scores.

	Lang. family →	Germanic			Romance			Slavic-Latin					Slavic-Cyrillic			Others				
	Generator (class)	de	en	nl	es	pt	ro	cs	hr	pl	sk	sl	bg	ru	uk	hu	el	ar	zh	all
Qwen3-4B-Base	Llama-2-70b-chat-hf	0.96	0.96	0.96	0.94	0.96	0.96	0.97	0.97	0.98	0.97	0.98	0.96	0.96	0.95	0.97	0.98	0.98	0.87	0.96
	Mistral-7B-Instruct-v0.2	0.96	0.95	0.98	0.97	0.97	0.98	0.97	0.98	0.98	0.97	0.96	0.97	0.95	0.92	0.95	0.95	0.98	0.83	0.96
	aya-101	0.94	**0.99**	0.97	0.97	0.97	0.98	0.98	0.96	0.99	0.99	0.99	0.99	0.98	0.99	0.99	0.98	0.98	**0.99**	0.98
	gpt-3.5-turbo-0125	0.98	0.96	0.99	0.98	**0.99**	**0.99**	0.99	**1.00**	**0.99**	**1.00**	0.99	0.99	0.97	0.98	0.99	**0.99**	0.99	0.93	0.99
	human	**0.99**	0.98	**0.99**	0.98	0.98	0.99	1.00	0.99	0.99	0.99	0.99	0.98	0.97	0.99	0.99	0.99	0.99	**0.99**	0.99
	opt-iml-max-30b	0.98	0.98	0.99	**0.99**	0.99	0.98	**1.00**	0.99	0.99	1.00	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	0.99	**1.00**	0.99	**0.99**
	v5-Eagle-7B-HF	0.91	0.92	0.95	0.92	0.92	0.95	0.97	0.93	0.88	0.96	0.94	0.92	0.88	0.90	0.88	0.92	0.95	0.83	0.92
	vicuna-13b	0.93	0.89	0.97	0.90	0.95	0.94	0.95	0.96	0.94	0.98	0.95	0.96	0.95	0.96	0.94	0.92	0.98	0.88	0.94
mdok	Llama-2-70b-chat-hf	0.96	0.98	0.97	0.96	0.96	0.98	0.97	0.97	0.96	0.97	0.98	0.96	0.95	0.96	0.97	0.98	0.96	0.86	0.96
	Mistral-7B-Instruct-v0.2	0.94	0.91	0.93	0.95	0.95	0.95	0.90	0.97	0.97	0.97	0.93	0.90	0.94	0.89	0.89	0.97	0.98	0.91	0.94
	aya-101	0.98	**0.99**	0.99	**0.99**	0.98	0.98	0.99	0.98	0.98	0.99	0.99	0.99	0.98	1.00	0.99	0.98	0.97	0.99	**0.99**
	gpt-3.5-turbo-0125	**0.99**	0.98	**1.00**	0.97	**0.99**	**1.00**	0.99	**1.00**	**0.99**	0.99	**1.00**	0.99	0.97	0.99	0.99	**0.99**	**0.99**	0.90	0.99
	human	0.97	0.98	0.99	0.97	0.98	0.99	0.99	0.99	0.98	0.99	0.99	0.97	0.96	0.98	0.96	0.95	0.98	0.99	0.98
	opt-iml-max-30b	0.93	0.96	0.98	0.94	0.97	0.96	**1.00**	0.99	0.98	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	0.99	0.99	**0.99**	0.98
	v5-Eagle-7B-HF	0.96	0.96	0.98	0.92	0.96	0.98	0.98	0.96	0.95	0.98	0.97	0.96	0.90	0.95	0.95	0.92	0.97	0.86	0.95
	vicuna-13b	0.91	0.87	0.93	0.90	0.91	0.93	0.96	0.96	0.94	0.99	0.95	0.94	0.93	0.94	0.92	0.98	0.98	0.92	0.94
OTBDetector	Llama-2-70b-chat-hf	0.96	0.97	0.97	0.96	0.95	0.98	0.98	0.98	0.98	0.97	0.98	0.97	0.98	0.98	0.95	0.98	0.98	0.87	0.97
	Mistral-7B-Instruct-v0.2	0.90	0.63	0.93	0.93	0.95	0.96	0.91	0.95	0.98	0.98	0.94	0.93	0.89	0.89	0.89	0.96	0.97	0.87	0.92
	aya-101	0.97	0.98	0.98	0.97	0.97	0.98	0.98	0.99	0.98	0.98	0.99	0.98	0.97	0.98	0.97	0.97	0.97	0.91	0.97
	gpt-3.5-turbo-0125	**0.99**	**0.98**	**1.00**	**0.99**	**0.99**	**0.99**	0.99	**0.99**	**0.99**	**1.00**	**1.00**	0.99	0.96	0.98	1.00	**0.99**	**1.00**	0.94	**0.99**
	human	0.97	0.92	0.99	0.96	0.96	0.99	0.99	0.98	0.98	0.99	0.99	0.98	0.97	0.97	0.98	0.95	0.98	**0.98**	0.98
	opt-iml-max-30b	0.90	0.87	0.97	0.93	0.94	0.97	**1.00**	0.98	0.98	0.99	0.99	**1.00**	**0.99**	**1.00**	**1.00**	0.95	0.98	0.98	0.97
	v5-Eagle-7B-HF	0.89	0.81	0.89	0.87	0.89	0.93	0.94	0.92	0.89	0.95	0.93	0.88	0.80	0.89	0.92	0.92	0.92	0.70	0.89
	vicuna-13b	0.88	0.74	0.88	0.74	0.85	0.92	0.91	0.93	0.88	0.97	0.93	0.95	0.87	0.91	0.89	0.93	0.98	0.78	0.89
XLM-R-large	Llama-2-70b-chat-hf	0.96	**0.99**	0.98	0.97	0.98	0.98	0.97	0.98	0.97	0.99	0.98	0.98	0.98	0.97	0.95	0.98	0.98	0.84	0.97
	Mistral-7B-Instruct-v0.2	0.84	0.45	0.83	0.84	0.90	0.92	0.82	0.83	0.97	0.91	0.83	0.75	0.67	0.70	0.82	0.91	0.92	0.71	0.82
	aya-101	0.97	0.98	0.99	0.97	0.98	0.99	0.98	0.98	0.98	0.99	0.99	0.99	0.97	0.99	0.99	0.97	0.98	0.90	0.98
	gpt-3.5-turbo-0125	**0.99**	0.98	**1.00**	**0.98**	**0.99**	**1.00**	**1.00**	**1.00**	**0.99**	**1.00**	**1.00**	0.99	0.97	**1.00**	**1.00**	**1.00**	**1.00**	0.96	**0.99**
	human	0.95	0.87	0.97	0.95	0.96	0.98	0.99	0.98	0.96	0.99	0.99	0.96	0.93	0.97	0.96	0.92	0.94	0.95	0.96
	opt-iml-max-30b	0.85	0.76	0.96	0.88	0.89	0.92	0.99	0.98	0.98	0.98	0.99	**1.00**	**0.99**	1.00	**1.00**	0.96	0.98	**0.97**	0.95
	v5-Eagle-7B-HF	0.76	0.62	0.77	0.69	0.73	0.82	0.90	0.89	0.84	0.88	0.92	0.84	0.70	0.83	0.87	0.85	0.84	0.57	0.80
	vicuna-13b	0.76	0.41	0.73	0.48	0.58	0.79	0.81	0.85	0.75	0.94	0.86	0.87	0.69	0.78	0.78	0.86	0.92	0.68	0.77
RoBERTa-large	Llama-2-70b-chat-hf	0.96	**0.99**	0.96	**0.97**	0.96	0.98	0.96	0.97	0.96	0.95	0.96	0.94	0.95	0.92	0.95	**0.95**	0.96	0.70	0.95
	Mistral-7B-Instruct-v0.2	0.85	0.62	0.83	0.84	0.92	0.90	0.77	0.89	0.97	0.81	0.82	0.56	0.58	0.52	0.77	0.65	0.77	0.57	0.77
	aya-101	0.88	0.97	0.93	0.87	0.92	0.93	0.96	0.94	0.93	0.94	0.96	0.93	0.89	0.95	0.94	0.90	0.92	**0.98**	0.93
	gpt-3.5-turbo-0125	**0.97**	0.97	0.95	0.95	0.96	0.97	0.95	0.97	0.92	0.97	0.97	0.79	0.78	0.81	0.86	0.72	0.69	0.71	0.89
	human	0.88	0.89	0.94	0.86	0.89	0.95	0.93	0.88	0.87	0.97	0.94	0.72	0.74	0.83	0.84	0.61	0.83	0.91	0.87
	opt-iml-max-30b	0.93	0.91	**0.98**	0.93	**0.96**	**0.98**	**0.99**	**0.99**	**0.97**	**0.99**	**0.99**	**0.99**	**0.97**	**0.98**	**0.96**	0.90	**0.97**	0.87	**0.96**
	v5-Eagle-7B-HF	0.72	0.68	0.73	0.72	0.73	0.76	0.79	0.83	0.72	0.80	0.79	0.41	0.49	0.61	0.71	0.38	0.53	0.41	0.67
	vicuna-13b	0.77	0.51	0.79	0.59	0.72	0.82	0.84	0.75	0.76	0.86	0.83	0.76	0.68	0.67	0.81	0.64	0.82	0.65	0.74
StatEnsemble	Llama-2-70b-chat-hf	0.73	**0.87**	0.87	0.81	0.83	0.80	0.71	0.80	0.79	0.48	0.37	0.69	0.67	0.67	0.66	0.36	0.20	0.38	0.68
	Mistral-7B-Instruct-v0.2	0.53	0.57	0.71	0.47	0.55	0.73	0.67	0.52	0.48	**0.90**	0.64	0.67	0.42	0.57	0.71	0.54	0.46	0.37	0.60
	aya-101	0.76	0.85	0.83	**0.86**	**0.91**	0.77	**0.78**	0.63	0.72	0.67	0.70	0.83	0.74	0.84	0.62	**0.92**	0.89	0.59	0.78
	gpt-3.5-turbo-0125	0.76	0.50	0.74	0.81	0.83	0.88	0.76	0.65	0.86	0.68	0.17	0.89	0.87	0.84	0.85	0.88	0.84	0.33	0.75
	human	0.74	0.24	0.76	0.51	0.61	**0.88**	0.77	**0.86**	**0.87**	0.86	0.75	0.80	0.76	0.83	0.81	0.74	0.70	0.81	0.75
	opt-iml-max-30b	**0.87**	0.13	**0.88**	0.59	0.52	0.28	0.58	0.65	0.84	0.74	**0.80**	**0.98**	**0.99**	**0.85**	**0.94**	0.87	**0.98**	**0.98**	**0.79**
	v5-Eagle-7B-HF	0.24	0.38	0.28	0.46	0.43	0.38	0.18	0.28	0.16	0.15	0.07	0.08	0.14	0.08	0.10	0.05	0.10	0.19	0.22
	vicuna-13b	0.53	0.37	0.48	0.39	0.37	0.40	0.27	0.32	0.54	0.17	0.24	0.31	0.44	0.37	0.51	0.23	0.22	0.32	0.37
Fast-DetectGPT	Llama-2-70b-chat-hf	**0.85**	**0.98**	**0.94**	**0.94**	**0.91**	**0.90**	**0.82**	**0.88**	**0.84**	0.44	**0.54**	0.74	0.76	0.83	0.79	0.08	0.13	0.42	**0.75**
	Mistral-7B-Instruct-v0.2	0.41	0.01	0.38	0.19	0.28	0.47	0.43	0.41	0.53	0.02	0.51	0.43	0.52	0.43	0.37	0.00	0.03	0.24	0.33
	aya-101	0.20	0.19	0.18	0.21	0.24	0.19	0.19	0.22	0.23	0.21	0.24	0.26	0.19	0.20	0.21	0.24	0.24	0.11	0.21
	gpt-3.5-turbo-0125	0.43	0.17	0.43	0.35	0.26	0.25	0.45	0.51	0.40	0.41	0.12	0.48	0.42	0.38	0.50	0.36	0.45	0.18	0.37
	human	0.41	0.38	0.38	0.40	0.41	0.40	0.41	0.40	0.37	0.39	0.39	0.47	0.45	0.47	0.45	0.41	0.40	0.19	0.40
	opt-iml-max-30b	0.57	0.01	0.63	0.21	0.31	0.17	0.28	0.13	0.64	**0.51**	0.51	**0.99**	**0.97**	**0.97**	**0.88**	**0.98**	**0.96**	**0.96**	0.67
	v5-Eagle-7B-HF	0.16	0.04	0.11	0.06	0.11	0.05	0.19	0.16	0.13	0.18	0.10	0.19	0.13	0.16	0.18	0.03	0.02	0.05	0.11
	vicuna-13b	0.16	0.01	0.09	0.10	0.09	0.12	0.14	0.10	0.20	0.14	0.20	0.18	0.21	0.19	0.21	0.05	0.04	0.16	0.13
Binoculars	Llama-2-70b-chat-hf	0.90	**1.00**	0.77	**0.98**	**0.96**	0.56	0.26	0.12	0.69	0.42	0.20	0.14	0.17	0.14	0.10	0.33	0.25	**0.76**	0.57
	Mistral-7B-Instruct-v0.2	0.25	0.00	0.40	0.06	0.22	0.51	0.52	0.61	0.57	0.47	0.62	0.25	0.56	0.35	0.59	0.55	0.65	0.36	0.44
	aya-101	0.10	0.15	0.12	0.06	0.10	0.17	0.13	0.11	0.14	0.08	0.18	0.03	0.13	0.03	0.17	0.10	0.07	0.06	0.11
	gpt-3.5-turbo-0125	0.03	0.00	0.05	0.00	0.00	0.06	0.03	0.01	0.02	0.01	0.05	0.01	0.05	0.01	0.01	0.03	0.02	0.01	0.02
	human	**0.90**	0.81	**0.90**	0.84	0.86	**0.97**	**0.98**	**0.96**	**0.93**	**0.96**	**0.95**	**0.95**	**0.80**	**0.90**	**0.97**	**0.81**	**0.82**	0.59	**0.89**
	opt-iml-max-30b	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	v5-Eagle-7B-HF	0.17	0.01	0.22	0.05	0.05	0.32	0.06	0.12	0.28	0.01	0.22	0.01	0.10	0.02	0.06	0.09	0.03	0.13	0.11
	vicuna-13b	0.32	0.01	0.34	0.04	0.16	0.52	0.20	0.25	0.38	0.10	0.20	0.02	0.15	0.06	0.14	0.21	0.06	0.27	0.20
	Writing script →	Lat	Lat	Lat	Lat	Lat	Lat	Lat	Lat	Lat	Lat	Lat	Cyr	Cyr	Cyr	Lat	Grk	Arab	Han	

Table 8: Per-generator multi-lingual performance (weighted F_1) of the selected methods using the MULTITUDE data.

		Lang. family →	Germanic			Romance			Slavic-Latin					Slavic-Cyrillic			Others				
		Generator (class)	de	en	nl	es	pt	ro	cs	hr	pl	sk	sl	bg	ru	uk	hu	el	ar	zh	all
en	mdok	Llama-2-70b-chat-hf	0.32	0.96	0.59	0.54	0.12	0.59	0.05	0.23	0.38	0.27	0.12	0.07	0.07	0.19	0.09	0.05	0.05	0.08	0.31
		Mistral-7B-Instruct-v0.2	0.63	0.90	0.88	0.76	0.16	0.50	0.41	0.31	0.05	0.09	0.31	0.03	0.03	0.19	0.46	0.05	0.01	0.04	0.38
		aya-101	0.67	**1.00**	0.79	0.72	0.34	0.76	0.28	0.14	0.27	0.04	0.16	0.04	0.18	0.06	0.33	0.00	0.00	0.05	0.39
		gpt-3.5-turbo-0125	0.42	0.97	0.87	0.87	0.40	0.92	0.09	0.29	0.24	0.29	0.04	0.00	0.42	0.06	0.16	0.00	0.03	0.05	0.42
		human	0.15	0.96	0.11	0.44	0.74	0.46	**0.98**	0.94	**0.95**	0.37	**0.96**	0.01	0.01	0.10	**0.86**	0.16	0.03	0.01	0.56
		opt-iml-max-30b	**0.81**	0.97	**0.89**	**0.99**	**0.99**	**0.93**	0.82	**0.95**	0.82	**0.97**	0.90	**0.93**	**0.87**	**0.90**	0.80	**0.91**	**0.87**	**0.65**	**0.89**
		v5-Eagle-7B-HF	0.71	0.91	0.63	0.82	0.84	0.92	0.65	0.51	0.87	0.42	0.26	0.44	0.60	0.67	0.25	0.25	0.84	0.65	0.66
		vicuna-13b	0.51	0.90	0.37	0.54	0.15	0.35	0.20	0.26	0.17	0.12	0.15	0.57	0.78	0.45	0.29	0.28	0.10	0.53	0.41
	OTBDetector	Llama-2-70b-chat-hf	0.07	0.90	0.12	0.16	0.07	0.12	0.03	0.05	0.09	0.10	0.01	0.03	0.09	0.03	0.01	0.02	0.03	0.04	0.14
		Mistral-7B-Instruct-v0.2	0.17	0.89	0.31	0.51	0.08	0.26	0.07	0.09	0.03	0.02	0.02	0.05	0.28	0.19	0.02	0.06	0.01	0.08	0.21
		aya-101	**0.94**	0.97	0.82	**0.89**	0.76	**0.84**	0.69	0.60	0.48	0.57	0.70	0.59	0.41	0.35	0.73	0.57	0.53	0.30	0.68
		gpt-3.5-turbo-0125	0.84	0.97	**0.85**	0.81	0.78	0.81	0.78	0.74	0.81	0.74	0.59	0.63	0.76	0.69	0.48	0.69	0.70	0.36	0.74
		human	0.45	**0.97**	0.76	0.40	0.77	0.39	0.82	0.86	0.81	0.72	0.79	0.60	0.43	0.49	0.82	0.35	0.27	0.38	0.65
		opt-iml-max-30b	0.60	0.87	0.82	0.84	**0.92**	0.78	**0.91**	**0.95**	**0.91**	**0.97**	**0.99**	**0.96**	**0.86**	**0.95**	**0.88**	**0.94**	**0.93**	**0.98**	**0.90**
		v5-Eagle-7B-HF	0.89	0.72	0.84	0.83	0.85	0.84	0.71	0.82	0.86	0.68	0.67	0.74	0.80	0.85	0.69	0.66	0.76	0.85	0.79
		vicuna-13b	0.55	0.90	0.60	0.85	0.47	0.82	0.23	0.35	0.45	0.10	0.13	0.19	0.65	0.44	0.14	0.07	0.06	0.10	0.45
es	mdok	Llama-2-70b-chat-hf	0.45	0.95	0.69	0.77	0.26	0.94	0.08	0.28	0.48	0.33	0.11	0.15	0.26	0.23	0.11	0.05	0.07	0.14	0.41
		Mistral-7B-Instruct-v0.2	0.91	0.73	0.95	**0.98**	0.91	0.96	0.76	0.66	0.86	0.35	0.66	0.11	0.19	0.19	0.66	0.24	0.03	0.27	0.65
		aya-101	0.87	**0.99**	**0.96**	0.96	0.83	**0.98**	0.57	0.67	0.81	0.57	0.72	0.07	0.61	0.47	0.80	0.05	0.05	0.17	0.69
		gpt-3.5-turbo-0125	0.18	0.57	0.85	0.93	0.78	0.97	0.33	0.40	0.61	0.22	0.05	0.03	0.55	0.10	0.10	0.00	0.16	0.00	0.45
		human	0.84	0.70	0.77	0.96	0.94	0.98	0.98	0.98	0.98	0.83	0.99	0.69	0.62	0.88	0.94	0.53	0.63	0.00	0.83
		opt-iml-max-30b	**0.97**	0.75	0.91	0.97	**0.99**	0.90	**0.99**	**1.00**	**0.98**	**0.99**	**0.99**	**0.97**	**0.93**	**0.96**	**0.99**	**0.97**	**1.00**	0.74	**0.95**
		v5-Eagle-7B-HF	0.82	0.93	0.71	0.89	0.93	0.90	0.77	0.66	0.92	0.58	0.59	0.48	0.72	0.72	0.42	0.21	0.63	**0.76**	0.73
		vicuna-13b	0.76	0.60	0.48	0.78	0.43	0.88	0.44	0.36	0.24	0.22	0.44	0.78	0.80	0.76	0.52	0.46	0.09	0.56	0.57
	OTBDetector	Llama-2-70b-chat-hf	0.55	**0.96**	0.54	0.84	0.49	0.94	0.25	0.38	0.47	0.28	0.08	0.29	0.71	0.62	0.07	0.05	0.08	0.17	0.49
		Mistral-7B-Instruct-v0.2	0.65	0.66	0.77	0.92	0.78	0.88	0.28	0.24	0.90	0.06	0.10	0.22	0.48	0.32	0.10	0.07	0.03	0.31	0.50
		aya-101	0.88	0.88	0.84	0.95	0.91	0.90	0.60	0.90	0.81	0.83	0.83	0.84	0.75	0.71	0.82	0.78	0.77	0.26	0.81
		gpt-3.5-turbo-0125	0.72	0.85	0.84	**0.96**	**0.96**	**0.96**	0.91	0.72	0.91	0.71	0.72	0.69	0.73	0.69	0.44	0.58	0.78	0.28	0.77
		human	0.83	0.78	0.94	0.85	0.89	0.89	0.92	0.97	0.96	**0.98**	0.94	0.81	0.75	0.83	0.95	0.78	0.87	0.93	0.89
		opt-iml-max-30b	**0.98**	0.55	**0.96**	0.93	0.95	0.91	**0.99**	**0.99**	**0.99**	0.98	**0.99**	**0.99**	**0.98**	**0.96**	**1.00**	**0.99**	**0.99**	**0.93**	**0.96**
		v5-Eagle-7B-HF	0.90	0.56	0.82	0.78	0.87	0.67	0.79	0.76	0.83	0.69	0.71	0.84	0.77	0.78	0.63	0.59	0.65	0.78	0.75
		vicuna-13b	0.47	0.56	0.38	0.79	0.40	0.79	0.20	0.20	0.33	0.07	0.08	0.08	0.42	0.26	0.09	0.08	0.05	0.07	0.33
ru	mdok	Llama-2-70b-chat-hf	0.89	**0.96**	0.93	0.92	0.75	**0.95**	0.73	0.72	0.89	0.46	0.81	0.80	0.94	0.67	0.57	0.06	0.07	0.17	0.73
		Mistral-7B-Instruct-v0.2	0.92	0.43	0.88	0.86	0.83	0.62	0.92	0.85	0.71	0.91	0.83	0.87	0.90	0.82	0.83	0.36	0.00	0.51	0.76
		aya-101	**0.97**	0.91	**0.98**	0.94	0.85	0.89	0.91	0.98	0.94	0.93	**0.99**	0.95	0.98	0.96	0.97	0.61	0.55	0.52	0.89
		gpt-3.5-turbo-0125	0.69	0.58	0.95	**0.94**	0.87	0.92	0.93	0.92	0.95	0.89	0.88	0.82	0.97	0.95	0.86	0.40	0.77	0.18	0.83
		human	0.08	0.38	0.72	0.69	**0.93**	0.55	**0.96**	0.97	**0.97**	0.93	0.97	0.97	0.96	0.96	0.88	0.77	0.77	0.00	0.80
		opt-iml-max-30b	0.91	0.45	0.68	0.80	0.75	0.73	0.84	**0.98**	0.89	**0.98**	0.97	**1.00**	**1.00**	**1.00**	**1.00**	**0.98**	**1.00**	**0.98**	**0.90**
		v5-Eagle-7B-HF	0.86	0.96	0.87	0.84	0.87	0.89	0.90	0.90	0.93	0.91	0.93	0.85	0.80	0.90	0.72	0.52	0.87	0.60	0.85
		vicuna-13b	0.78	0.33	0.65	0.65	0.72	0.45	0.56	0.72	0.54	0.66	0.58	0.79	0.93	0.79	0.62	0.49	0.29	0.72	0.65
	OTBDetector	Llama-2-70b-chat-hf	0.85	**0.96**	0.84	**0.91**	**0.89**	**0.88**	0.77	0.74	0.89	0.46	0.50	0.79	0.96	0.94	0.31	0.08	0.10	0.37	0.73
		Mistral-7B-Instruct-v0.2	0.57	0.43	0.55	0.33	0.68	0.21	0.73	0.80	0.68	0.95	0.69	0.78	0.88	0.88	0.43	0.13	0.08	0.62	0.62
		aya-101	**0.85**	0.70	0.88	0.80	0.83	0.68	0.84	0.96	0.88	0.93	0.92	0.92	0.89	0.90	0.94	0.86	0.81	0.17	0.84
		gpt-3.5-turbo-0125	0.66	0.57	0.93	0.72	0.80	0.73	**0.90**	0.93	**0.95**	0.91	0.93	0.93	0.95	0.96	0.96	0.89	0.91	0.65	**0.86**
		human	0.43	0.76	**0.96**	0.52	0.68	0.74	0.87	**0.98**	0.94	**0.98**	**0.96**	0.89	0.92	0.90	0.97	0.78	0.65	0.84	0.84
		opt-iml-max-30b	0.43	0.05	0.26	0.22	0.22	0.10	0.77	0.92	0.79	0.89	0.87	**1.00**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.97**	0.77
		v5-Eagle-7B-HF	0.80	0.60	0.79	0.67	0.50	0.61	0.82	0.87	0.74	0.86	0.92	0.85	0.67	0.86	0.89	0.89	0.88	0.41	0.77
		vicuna-13b	0.82	0.43	0.80	0.72	0.85	0.65	0.82	0.80	0.79	0.60	0.69	0.74	0.80	0.79	0.49	0.28	0.15	0.61	0.68
en-es-ru	mdok	Llama-2-70b-chat-hf	0.71	0.93	0.82	0.88	0.62	0.93	0.56	0.48	0.82	0.44	0.49	0.68	0.92	0.70	0.28	0.05	0.07	0.15	0.64
		Mistral-7B-Instruct-v0.2	0.96	0.90	0.94	**0.98**	0.97	0.96	0.93	0.87	0.94	0.79	0.88	0.79	0.84	0.83	0.78	0.32	0.09	0.31	0.82
		aya-101	**0.98**	**1.00**	**0.99**	0.97	0.95	0.98	0.91	0.97	0.97	0.90	0.97	0.94	0.99	0.98	0.97	0.50	0.66	0.55	0.92
		gpt-3.5-turbo-0125	0.93	0.96	0.96	0.95	0.93	0.95	0.92	0.92	0.94	0.89	0.86	0.89	0.95	0.91	0.83	0.60	0.87	0.26	0.88
		human	0.86	0.96	0.92	0.96	**0.97**	**0.99**	**0.98**	0.99	**0.97**	0.95	0.98	0.95	0.95	0.94	0.95	0.81	0.87	0.23	0.92
		opt-iml-max-30b	0.90	0.94	0.76	0.97	0.94	0.90	0.97	**1.00**	0.95	**0.99**	**0.99**	**1.00**	**1.00**	**0.99**	**0.99**	**0.99**	**1.00**	**0.95**	**0.96**
		v5-Eagle-7B-HF	0.88	0.89	0.82	0.85	0.92	0.87	0.85	0.85	0.94	0.79	0.84	0.80	0.81	0.83	0.65	0.48	0.80	0.80	0.82
		vicuna-13b	0.79	0.89	0.67	0.83	0.62	0.83	0.51	0.67	0.48	0.48	0.54	0.78	0.88	0.75	0.61	0.48	0.19	0.53	0.66
	OTBDetector	Llama-2-70b-chat-hf	0.54	0.93	0.58	0.81	0.67	0.88	0.55	0.58	0.72	0.35	0.23	0.56	0.91	0.84	0.12	0.05	0.06	0.12	0.58
		Mistral-7B-Instruct-v0.2	0.65	0.72	0.61	0.85	0.80	0.87	0.58	0.45	0.74	0.26	0.39	0.59	0.85	0.77	0.14	0.07	0.02	0.40	0.59
		aya-101	0.95	0.95	0.90	0.93	0.84	0.91	0.62	0.94	0.81	0.89	0.88	0.90	0.89	0.89	0.89	0.77	0.80	0.24	0.85
		gpt-3.5-turbo-0125	0.94	**0.99**	**0.96**	**0.95**	**0.97**	**0.96**	0.98	0.98	0.97	0.95	0.98	0.97	0.96	0.97	0.97	0.89	0.95	0.55	0.95
		human	0.74	0.90	0.96	0.81	0.91	0.88	0.93	0.96	0.94	0.98	0.95	0.87	0.86	0.86	0.97	0.82	0.71	0.63	0.88
		opt-iml-max-30b	**0.95**	0.79	0.88	0.90	0.92	0.87	**0.99**	**0.99**	**0.99**	**0.98**	**0.99**	**1.00**	**0.99**	**0.99**	**1.00**	**0.99**	**0.99**	**0.99**	**0.96**
		v5-Eagle-7B-HF	0.84	0.50	0.80	0.66	0.63	0.59	0.82	0.81	0.71	0.70	0.82	0.75	0.57	0.77	0.67	0.58	0.69	0.66	0.70
		vicuna-13b	0.76	0.66	0.75	0.83	0.71	0.76	0.52	0.64	0.64	0.33	0.35	0.51	0.78	0.72	0.31	0.20	0.12	0.37	0.58
		Writing script →	Lat	Lat	Lat	Lat	Lat	Lat	Lat	Lat	Lat	Lat	Lat	Cyr	Cyr	Cyr	Lat	Grk	Arab	Han	

Table 9: Per-generator cross-lingual performance (weighted F_1) of two of the best AA methods using the MULTITUDE data.