AI 'News' Content Farms Are Easy to Make and Hard to Detect: A Case Study in Italian

Anonymous ACL submission

Abstract

Large Language Models (LLMs) are increasingly used as 'content farm' Models (CFMs), to generate synthetic text that could pass for real news articles. This is already happening even for languages that do not have high-quality monolingual LLMs. We show that fine-tuning Llama LLM, mostly trained on English, on as little as 40K Italian news articles, is sufficient for producing news-like texts that native speakers of Italian struggle to identify as synthetic.

We investigate three LLMs and three methods of detecting synthetic texts (log-likelihood, DetectGPT, and supervised classification), finding that they all perform better than human raters, but they are all impractical in the real world (requiring either access to token likelihood information or a large dataset of CFM texts). We also explore the possibility of creating a proxy CFM: an LLM fine-tuned on a similar dataset to one used by the real 'content farm'. We find that even a small amount of fine-tuning data suffices for creating a successful detector, but we need to know which base LLM is used, which is a major challenge.

Our results suggest that there are currently no practical methods for detecting synthetic news 'in the wild', while generating them is too easy. We highlight the urgency of more NLP research on this problem.

1 Introduction

The modern Large Language Models (LLMs) can generate increasingly fluent and plausible-sounding texts, which sparks concerns about their potential misuse by bad actors. One of the emerging problems is AI-driven news "content farms": news-like sites filled with synthetic texts that are not necessarily serving a misinformation campaign, but are plausible-looking enough to deceive the readers and generate web traffic. Already in May 2023 NewsGuard reported that they found 49 such 'outlets' (Sadeghi and Arvanitis, 2023), and on February 12th 2024 their count[1] was 702. Sometimes such sites publish original 'content', and sometimes they automatically 'rewrite' articles from real news outlets without attribution (Brewster et al., 2023). They are primarily created for serving programmatic ads, and even major brands may unwittingly support such 'outlets' (Ryan-Mosley, 2023). They already operate in many languages.[2] The problem will likely only get worse with time, and it needs more attention both on the policy & regulation side and from the NLP researchers.

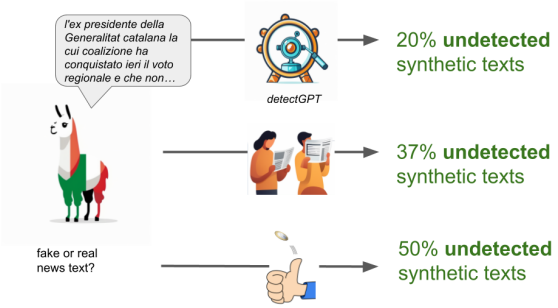


Figure 1: Detecting synthetic Italian news text generated by fine-tuned Llama-65B: error rates for DetectGPT, native speakers of Italian and random guess.

We illustrate how easy it is to create 'content farm' models (CFMs), and how hard to detect them, by considering a case that should be relatively tricky. In §3 we successfully turn Llama, a relatively old LLM mostly trained on English, into an Italian news CFM – by fine-tuning it on only 40k Italian news texts. For Llama65B this turns out to be sufficient to mislead native speakers of Italian, who identify synthetic texts with only 64% accuracy, vs 50% random guess (see §4). We also find that existing detection methods perform better

[1] https://www.newsguardtech.com/special-reports/ai-tracking-center/

[2] NewsGuard AI Tracking Center currently states that they found such 'outlets' in Arabic, Chinese, Czech, Dutch, English, French, German, Indonesian, Italian, Korean, Portuguese, Spanish, Tagalog, Thai, and Turkish.

than humans (§5), but are impractical in the real world (§7).

In §6 we explore an alternative approach: fine-tuning another LLM as a proxy for the real CFM and relying on its token likelihood scores as proxies for the scores of the real CFM. We find that this works with even a small amount of fine-tuning data (only 3% of the full fine-tuning dataset) but only given that we know which base LLM was used as CFM. We also experiment with using these proxy scores to identify the base LLM, but this method would also be impractical if there are dozens, if not hundreds possible alternatives.

We hope that our findings will boost similar investigations for other languages, and highlight the urgency of developing model-agnostic methods for synthetic text detection. To facilitate future research on Italian, we release *(i)* 15k news passages generated by models fine-tuned on the CHANGE-it dataset, an Italian news dataset (Mattei et al., 2020), *(ii)* ratings produced by 5 human annotators on 400 texts, with a balanced distribution of 50% human-written and 50% synthetic passages, and *(iii)* 600k synthetic alterations of both the original samples from the CHANGE-it dataset and the synthetic texts generated. We will not publicly release our fine-tuned LLMs (since they are best suited to be used as CFMs), but we welcome direct requests from researchers working on this problem.

2 Related work

Driven by the increasing number of strong openly available LLMs (Touvron et al., 2023a; Brown et al., 2020; Raffel et al., 2020; Jiang et al., 2023), several studies focused on the detection of synthetic text detection. Ghosal et al. (2023) identified two main groups of approaches: those based on token likelihood, and supervised classification. Among the former, Mitchell et al. (2023); Su et al. (2023); Hans et al. (2024) proposed detection methods that rely on language models token distribution. In terms of supervised detection, Verma et al. (2023) proposed an approach that relies on the availability of labelled datasets of synthetic and human-written texts. Wang et al. (2023) contributed a large multilingual, multi-domain dataset that can be used to benchmark synthetic text detection systems.

After the release of ChatGPT (OpenAI, 2022), a lot of studies (mostly not yet peer reviewed) focused specifically on the detection of text generated by ChatGPT and GPT-4 (OpenAI, 2023) (Dhaini et al., 2023). Since it is not possible to access token probabilities for a candidate generated text, some of this work relies on a proxy model (Vasilatos et al., 2023), but most rely on supervised classification (Mitrović et al., 2023; Liao et al., 2023; Guo et al., 2023; Liu et al., 2023), including OpenAI itself (Kirchner et al., 2023). Sometimes classical machine learning techniques are reported to perform well: e.g. Desaire et al. (2023) report a high performance on chemistry articles with XGBoost classifier and 20 features extracted from paragraphs.

The above work focuses on the scenario where the effort of detecting synthetic texts is on the user side. The complementary direction on the developer side is watermarking: ensuring that the LLM output creates some kind of statistical "signature" that would help to identify it. There are multiple proposals for how to do this (Fernandez et al., 2023; Kirchner et al., 2023; Kuditipudi et al., 2023; Li et al., 2023; Takezawa et al., 2023; Wu et al., 2023; Yoo et al., 2023, inter alia), and some initial results suggesting that such techniques could be sufficiently robust to human and machine paraphrasing (Kirchenbauer et al., 2023). However, most of the current 'open' LLMs are freely available without any watermarking, and as we will show, they are already sufficient to be used as CFMs.

For the problem of detecting synthetic text without watermarking, the current research focuses on either monolingual or multilingual-by-design LLMs, and most studies do not focus on a specific domain. We stress the importance of also investigating fine-tuned models because LLMs are expensive to both train and run inference on (Samsi et al., 2023). Hence, starting from a public, relatively small, but high-quality LLM and fine-tuning it for a specific type of content is probably the most plausible scenario for 'content farms' that aim to produce texts cheaply. To the best of our knowledge, this is the first study to focus on the scenario where the 'CFM' is fine-tuned for news, and in a language that it was not meant for originally.

3 Fine-tuning of Llama as an Italian 'Content Farm' Model

As our 'CFM', we choose the original Llama model (Touvron et al., 2023a), in 7B and 65B parameter versions. Since this is one of the first public high-performing LLMs of this size, our Llama results serve as a lower bound for what could be expected for Italian from later LLMs, such as Llama 2 (Tou-

vron et al., 2023b), Mistral (Jiang et al., 2023), and others. We do *not* suggest that it is possible to obtain good results with any language and any LLM (see also §9): such a transfer depends on the similarity between languages and their coverage in the training data. But given that 'content farms' have already been identified in at least 15 languages (see §1), many others could follow.

Our choice of Italian enables a lower-bound estimate of what could be expected of monolingual or multilingual LLMs with more exposure to the target language. Llama is a 'mostly-English' model, not intended to be multilingual. However, it was exposed to at least Italian Wikipedia (Touvron et al., 2023a, p.2), and the other resources it was trained on, such as C4 corpus (Raffel et al., 2020), made deliberate efforts to filter out non-English text. Hence, Wikipedia is likely the main, if not the only source of Italian in Llama.

We experiment with the original Llama baseline (65B pre-trained model with no extra training), and two versions of our Llama fine-tuned on Italian news, after 20K and 60K steps. The technical details for fine-tuning are provided in App. B.1. We remark that creating such a CFM now comes with very few technical or financial difficulties,[3] which could increase the number of bad actors.

4 Detection of Synthetic News in Italian By Native Speakers

Methodology. To assess whether native Italian speakers would be able to identify synthetic news texts generated by our CFM, we set up a crowd-based study following the general recommendations for human evaluation of automatically generated texts proposed by van der Lee et al. (2021). Specifically, we created 4 surveys, each comprising 100 questions. To maintain the rater engagement, we administered the surveys in five sessions, each comprising 20 questions. The raters were anonymously recruited among Italian native speakers via the Prolific[4] online crowd-sourcing platform. Five different raters participated in each survey session, with no limitations on the number of sessions a rater could undertake. The 20-question sessions took 8 mins 23 secs on average, and the raters were compensated at 9,68$ per hour.[5] The study involved a total of 93 distinct raters, with an average age of 32.01 ($\pm$10.76).

Model	Accuracy	STD	Fleiss k
Llama 7B pretrain	83.2	7.0	36.45
Llama 7B finetuned	69.5	12.2	22.30
Llama 65B pretrain	73.7	5.8	33.01
Llama 65B finetuned	64.2	11.2	20.56

Table 1: Accuracy and standard deviation achieved by human raters in assessing human-written versus machine-generated news. We report the inter-rater agreement measured as group Fleiss' k.

Each of the 4 surveys is aimed at assessing the texts generated by a different model: Llama 7B and 65B, both with and without fine-tuning on Italian news. Each question required raters to read two texts, denoted as A and B, and answer the question *"Text B follows text A, do you think text B is written by a machine?"*. The answer was a rating on a 5-point Likert scale, where 1 indicates 'certainly human-written' and 5 'certainly machine-generated'. In 50% of the questions, both text A and B were human-written news articles coming from the original CHANGE-it test dataset.

The topics included daily national political events (e.g. politicians' declarations), general news (e.g. climate catastrophes), and relevant international news (e.g. European leadership meetings).

To estimate the accuracy of human raters on this task, we map the scores on the 5-point scale to a binary score by computing the average score assigned by the raters to a given sample. We interpret the average score above 3 as indicating that a given sample B was generally perceived as machine-generated. We also experimented with a different thresholding approach, using the average rating as a threshold, that showed similar results; these results are available in App. C.

Results. Table 1 shows the outcome of our analysis. Since the raters' accuracy in detecting news generated by the largest fine-tuned Llama 65B is as low as 64%, we can answer our research question positively: **Llama can be fine-tuned to generate**

[3] Renting GPUs on cloud providers such as Amazon is now relatively cheap (approx 15$ per hour for 8x40Gb A100) and would require as little as 100$ to replicate one of our LLM training sessions and data generation. The technical barrier to fine-tuning LLMs is also low now, thanks to tools like Huggingface's Autotrain:`https://huggingface.co/autotrain`. We do not criticize open-sourcing such tools, but we hope that our results would highlight the necessity of more research on synthetic text detection.

[4] `https://www.prolific.co/`. The participant group was balanced in terms of gender (49.46% female) and student status (50.60% reported being students).

[5] This hourly payment rate was certified as 'Fair' by the Prolific platform.

Prompt: *"[...] l'ex presidente della Generalitat catalana la cui coalizione ha conquistato ieri il voto regionale e che non..."* EN: [...] the former president of Catalan Generalitat, whose coalition won the regional election yesterday, and who...

Pre-trained: *... vuole rinunciare alla secessione. In the 6-week period prior to 12/06/19...* EN: ... does not want to give up the secession. In the 6-week period prior to 12/06/19 ...

Fine-tuned: *...aveva perso tempo per dire la sua. Da Bruxelles, dove si trova da allora ...* EN: ... does not waste time to mention his opinion. From Brussels, where he resides since...

Figure 2: Example: without fine-tuning on Italian, Llama is prone to switching to English.

hard-to-detect news-like text in Italian. We only used 40K samples for fine-tuning a relatively old model, so even more plausible-sounding synthetic text could likely be created with more data and more recent LLMs.

Overall, the raters' accuracy reflects two foreseeable trends: the smaller 7B models are easier to detect, and the fine-tuned models are harder to detect. Interestingly, the small 7B version, when fine-tuned on Italian, is identified by raters with accuracy close to the larger pretrained 65B.

Table 1 also reports the average inter-rater agreement for each survey. Fleiss κ (Fleiss et al., 1971) is in the range between 22%-36%, indicating a "moderate" agreement (Landis and Koch, 1977) consistent with similar human-evaluation studies (van der Lee et al., 2021). The raters agree more strongly when assessing non-fine-tuned models, possibly because they occasionally switch from Italian to English mid-generation, a characteristic identified by raters as indicative of machine-generated texts. Figure 2 shows an example of such a switch.

The qualitative analysis of 100 machine-generated instances (25 per model) showed that 46/100 examples had no obvious issues with language, but, as expected for LLM-generated texts, their content was factually incorrect and, worryingly, some factual errors were not obvious without extra fact-checking effort. Additionally, in this sample of 100 texts, we found 7 examples where the generated text contradicted the prompt, 8 cases of language-switching, 18 samples with grammatical errors and 21 with expressions that are grammatically correct but unnatural in Italian. Annotated examples for each model are shown in App. D.

We stress that this study focuses only on the problem of synthetic news-like text, which is sufficiently plausible for the 'content farms' to lead the users to 'news' websites and be served ads. Their success also likely depends on the quality of the 'headlines', their match to the interests of the audience, the position in the search engine rankings and other factors beyond the scope of this study. Still, having such texts is a necessary, though not sufficient condition for operating a 'content farm'.

5 Automatic Detection of Synthetic News in Italian

5.1 Approaches based on token likelihood

Methodology. Similarly to Jawahar et al. (2020); Sadasivan et al. (2023); Chakraborty et al. (2023), we attempt the zero-shot detection of artificially generated text. We experiment with two approaches for synthetic text detection. Both assume to have access to the likelihood of each token in a sentence, according to the model whose "authorship" is under analysis.

The *log-likelihood* measures how likely a sentence is according to the probability assigned by the model to each token. The core idea behind the *DetectGPT* score (Mitchell et al., 2023) is getting a more robust score by normalising the *log-likelihood* of a sentence based on modifications[6] of that same sentence generated by a different model (which we refer to as the *bootstrap model*).

Given a sentence, both *log-likelihood* and *DetectGPT* can be used with a threshold to tell if that sentence is more likely written by a human or by a language model. The threshold is estimated empirically based on a collection of synthetic and real samples. We refer the reader to Mitchell et al. (2023) for details of likelihood-based methods.

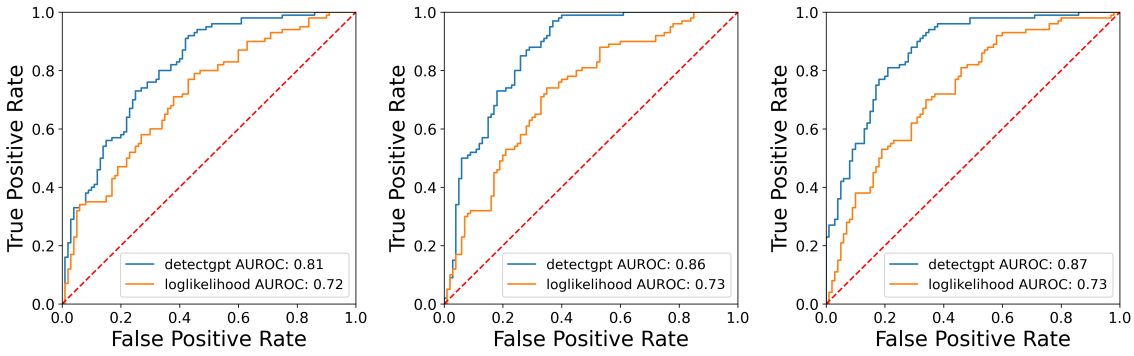
We start by validating our codebase and implementation by repeating prior experiments on xsum (Narayan et al., 2018) dataset in English, and we obtain comparable results (see App. E).

To measure the ROC of both *DetectGPT* and *log-likelihood* we select 100 random sentences and generate[7] another 100 by using the first 30 tokens of each sentence as a prompt for the model under analysis. After this, we clip all sentences to 150 tokens,[8] measure the score for each sample, and

[6] In our case, we are interested in the likelihood of a synthetic sentence estimated by Llama, vs the normalized likelihood that Llama assigns to modifications of that same sentence that we generate with T5 (Raffel et al., 2020) as the bootstrap model. See App. D for examples of such modifications.

[7] All generation tasks are performed using Nucleus Sampling for decoding (Holtzman et al., 2020).

[8] Due to different tokenizers, this step results in sentences with a varying number of words, but a similar length.



(a) Original Llama 65B (b) Fine-tuned epoch 2 (c) Fine-tuned epoch 6

Figure 3: ROC curve for *DetectGPT* and *log-likelihood*. In (a) for Llama 65B measured over 100 sentences from the CHANGE-it data-set (Italian), in (b) the same measure for Llama 65B model after 20,000 fine tuning steps on CHANGE-it training set and in (c) after 60,000 fine-tuning steps.

compute the AUROC on all 200 sentences (half human-written and half machine-completed).

To compare the detection approaches for the original and fine-tuned Llama, we use the CHANGE-it test dataset for Italian. As the bootstrap model, we use *IT5-large* (Sarti and Nissim, 2022) for Italian.

Results. Figure 3 shows the AUROC of *DetectGPT* and *log-likelihood* for all our models. *log-likelihood* does not seem to react to fine-tuning at all: the fine-tuned models have almost the same AUROC as the pre-trained one. *DetectGPT* does have a 5-6 points higher AUROC for the fine-tuned models. However, by qualitative analysis, see §4, we find that fine-tuning should have made the task more rather than less difficult. Without fine-tuning, our Llama CFM has a tendency to switch to English mid-generation, as we also observed this in the human detection study (see Figure 2, more examples available in App. D). This language switch should be a clear marker for detecting both the original 7B and the 65B models, and it vanishes after fine-tuning. But both log-likelihood and DetectGPT are missing this clear signal.

Although the *DetectGPT* and *log-likelihood* perform relatively well in our tests, we stress that this indicates a measure of the difficulty of this task, rather than a solution to synthetic news detection (see section 7). We remark that DetectGPT score can be turned into an accuracy measurement by fixing a threshold, for a direct comparison with the human evaluation accuracy. In our case, DetectGPT accuracy is $\approx 80\%$ using the median[9] score as the threshold for fine-tuned Llama 65B.

5.2 Supervised Detection of Synthetic Texts

A different approach to identifying synthetic texts is to train supervised classifiers (Liu et al., 2019; Conneau et al., 2020). However, this requires a labelled and balanced dataset of human and synthetic texts. To understand the challenges of this scenario, we use DICE, a different Italian news dataset focusing on crime news (Bonisoli et al., 2023). We mix DICE with the CHANGE-it data as well as the synthetic texts. This simulates a more realistic data collection process compared to previous studies that solely focused on fully in-domain data.

Methodology. To create synthetic news, we fine-tune the following recent models: llama2-7b, llama2-13b (Touvron et al., 2023b) and Mistral-7b (Jiang et al., 2023) on the full CHANGE-it training set (see App. B.2 for the fine-tuning details). We refer to the resulting fine-tuned models as *llama-2-7b_it*, *llama-2-13b_it*, and *mistral_it*.

For each of these models, we create a suite of datasets with training sets composed of 2K, 4K or 8K samples. In all settings, the test sets comprise 2K samples, namely 1K texts from the CHANGE-it test set and 1K synthetic news with the same titles. The training sets are built in two ways. In *in domain* setting, 50% texts are synthetic, and 50% are articles sourced from CHANGE-it. In *mixed source*, the human-authored articles come from two different datasets, 25% each (CHANGE-it and DICE). *Mixed source* is closer to the scenario where we do not know what dataset was used to fine-tune the CFM, and sample from a wide range of possible news articles. The more diverse this set of non-

[9] We choose this threshold knowing that the dataset is balanced and that DetectGPT is monotonic, otherwise we would need to tune it.

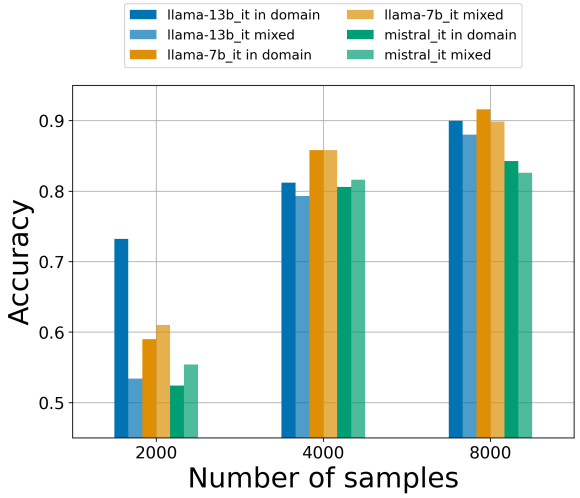


Figure 4: Accuracy of classifier based on xlm-RoBERTa-large for human/synthetic text classification task, for synthetic texts generated by three LLMs fine-tuned on CHANGE-it. The classifier was trained on 50% synthetic texts and either 50% CHANGE-it texts (*in domain*), or 25% texts from CHANGE-it and 25% from DICE (*mixed source*). Classification is only successful at at least 4K labeled samples, and the *mixed source* scenario is consistently more challenging.

synthetic examples, the harder the classification task will probably become.

Similarly to Wang et al. (2023), we experiment with two classifiers, based on RoBERTa (Liu et al., 2019) and XLM-RoBERTa (Conneau et al., 2020). The former is pre-trained on English and the latter specifically made for multilingual settings. We use the same hyper-parameters for both classifiers, and we train them for 3 epochs. The technical details are provided in App. B.3.

Results. Figure 4 shows that at 8k training samples *xlm-roberta-large* accuracy ranges between 84% and 92% depending on the generator. With 4k training samples, the accuracy decreases between 81% and 86%, with a similar difference between *in domain* and *mixed source*. With 2k training samples, the performance for all but 3 models drops to under 56% – i.e. almost random chance level. The RoBERTa-based classifier yields a similar trend. See App. B.3 for the graph for RoBERTa and numerical results for both experiments.

We find that for these datasets there is a threshold between 2k and 4k samples that marks a strong difference in the performance of the supervised classifier (20-25% loss in accuracy), while increasing the dataset size further provides diminishing returns (2-8% gain). Our *mixed source* scenario at 8K samples is also consistently harder for the classifier; at 4K the results are mixed across different LLMs, and at 2K it is hard to draw firm conclusion for *llama-2-7b_it* and *mistral_it* because their performance is generally low. This is with a multilingual classifier base; for RoBERTa-based classifier, the *mixed* setting is consistently harder in all settings (see App. B.3).

6 Detecting CFMs with Proxy Models

6.1 How Much Fine-tuning Do We Need?

So far we only considered the scenarios where the model generating synthetic texts is the same that computes the likelihood used to detect them. Let us now consider an alternative: a *proxy model* approximating the likelihood scores of the CFM. This could be expected to work if the proxy model and the CFM were fine-tuned on the same dataset. But that assumption is also too strong to be practical.

We explore relaxing it further to the scenario where we only have access to a small set of texts similar to the fine-tuned model outputs. This could be a small subset of the fine-tuning dataset or, more generally, samples from a similar distribution (e.g. different samples from a given newspaper or a social media account). We ask whether this is enough: **can we train a model on a small set of in-distribution texts, so that the likelihood it assigns to tokens is sufficient to detect texts from a fully fine-tuned CFM?**

Methodology. To answer this question, we fine-tune the three LLMs used in the supervised detection experiment in §5.2 (namely *llama-2-7b_it*, *llama-2-13b_it* and *mistral_it*) on varying subsets of the CHANGE-it dataset. We then use their likelihood and DetectGPT score to detect synthetic text from models fine-tuned on the full CHANGE-it dataset, as described in §5.

For that, we select 5000 samples from the CHANGE-it test set, and for each sample, we prompt all models with the title and initial tokens of the original article (see Appendix F for more details). This results in 3 datasets with 10k samples in each, 5k human written (from the CHANGE-it test set) and 5k synthetically generated news with the same titles, which can be used as a benchmark for the detection of synthetic news in Italian.

Afterwards, we select two subsets of the CHANGE-it training dataset with 3981 and 7862 samples, respectively 3% and 6% times the original fine-tuning dataset size. We fine-tune the same

Detector model	Generator model					
	llama-2-13b_it		*llama-2-7b_it*		*mistral_it*	
	dGPT	llh	dGPT	llh	dGPT	llh
llama-2-13b	0.73	0.61	0.54	0.40	0.56	0.43
llama-2-13b_it_3981	0.84	0.69	0.53	0.35	0.56	0.42
llama-2-13b_it_7862	0.85	0.70	0.53	0.34	0.56	0.41
llama-2-13b_it	0.87	0.70	0.48	0.27	0.55	0.39
llama-2-7b	0.58	0.49	0.75	0.59	0.57	0.46
llama-2-7b_it_3981	0.63	0.48	0.86	0.67	0.60	0.45
llama-2-7b_it_7862	0.63	0.47	0.87	0.68	0.60	0.44
llama-2-7b_it	0.62	0.44	0.88	0.66	0.61	0.44
mistral	0.54	0.46	0.52	0.40	0.68	0.54
mistral_it_3981	0.54	0.42	0.48	0.34	0.80	0.65
mistral_it_7862	0.54	0.41	0.47	0.32	0.81	0.67
mistral_it	0.44	0.29	0.35	0.20	0.94	0.85

Table 2: The AUROC achieved by all the models (rows) at different levels of fine-tuning, from pretrained only to fine-tuned on the full dataset. In all settings, the AUROC for models fine-tuned on 3981 and 7861 samples is very close to the results of the fully fine-tuned model. However, the best results are always on the diagonal cells, where the detector and generator models are the same.

LLMs on both of these smaller training sets, see App. B.2 for details on the fine-tuning procedure and Table 4 for a summary of the model naming.

Finally, for each of the three synthetic datasets generated with *llama-2-7b_it*, *llama-2-13b_it* and *mistral_it*, we compute *log-likelihood* and *Detect-GPT* score with all 12 models: 3 fine-tuned on the full CHANGE-it dataset, 3 fine-tuned on 3981 samples, 3 on 7862 samples, and the original LLMs without any fine-tuning.

Results. Table 2 suggests that the answer to our research question is positive: **a small subset of fine-tuning samples is indeed sufficient to detect a full fine-tuned model**. A model fine-tuned on only 3% of the fine-tuning dataset achieves between 86.1% and 95.4% of the AUROC measured for the fine-tuned one. However, this is only effective if the same LLM is both the generator and the detector, see App. G for the ROC curves in this case. Simply fine-tuning different LLMs does not make them similar enough to use one for detecting another. In the case of *mistral_it*, it actually gets worse after fine-tuning (we hypothesize that it could be due to differences in tokenization).

While this study focuses on the Italian news, in App. E.2 we perform the same experiments on the XSUM dataset in English, and come to equivalent conclusions: we can detect a fully fine-tuned model with a model that is fine-tuned on a small subset of the whole fine-tuning dataset.

6.2 Can Ensembling Help?

As shown in Table 2, fine-tuning on few samples is sufficient to achieve strong AUROC both with *log-likelihood* and *DetectGPT* on fine-tuned versions of the same LLM – but detecting fine-tuned versions of different LLMs is harder, and longer fine-tuning does not improve the performance of statistical detection methods. This makes the proxy model approach impractical, since in the real world we would not know which base LLM was used.

To address this limitation, a straightforward approach would be ensembling the *DetectGPT* score, which different candidate LLMs assign to a given sample. If texts from a certain source consistently get a high AUROC, this would signal that they are probably synthetic. This is reasonable on the assumption that the CFM is one of the recent high-performing open-source LLMs, and there are not too many of these. Still, to evaluate the potential of this approach we experiment with computing both the mean and max *DetectGPT* score among three LLMs with the same fine-tuning, to understand if this provides a score with higher AUROC.

The results are shown in Table 3. Neither the average nor the maximum *DetectGPT* score offer a simple solution, but in most cases, one of them yields a significantly higher AUROC than the random guess. We believe that this is overall a promising direction for developing new statistic approaches based on mixing the likelihoods of several models, but, once again, it would be difficult to scale to dozens of candidate LLMs.

CHANGE-it samples	Mode	*llama-2-7b_it*	*llama-2-13b_it*	*mistral_it*
Full	max	0.75	0.74	0.92
	mean	0.61	0.69	0.79
	random	0.57	0.65	0.68
3981	max	0.62	0.84	0.62
	mean	0.66	0.73	0.68
	random	0.63	0.63	0.64
7962	max	0.58	0.83	0.68
	mean	0.65	0.73	0.69
	random	0.62	0.63	0.64

Table 3: We experiment with ensembling the *DetectGPT* score measured by models with the same amount of fine-tuning (indicated in the first column). We devise three new scores: *random* is computed by randomly picking the *DetectGPT* score from one of the models, *mean* is computed by taking the average value of all models and *max* by taking the highest value.

7 Discussion: Are There Practical Solutions to Synthetic Text Detection?

As described in §1, the 'content farms' for news are already wide-spread, and it is in the public interest to identify such 'outlets', as soon as possible.

One key problem is that we would not know which base LLMs were used, and we would not have access to their token likelihood information. This completely precludes using methods like DetectGPT (§5.1). Our proxy model approach (§6) could address the latter problem, but the former is far from being solved (§6.2), and will only get harder as more LLMs are published. The offending CFM could also be a 'closed' model like GPT-4, to which we would never have white-box access.

Supervised approaches can be used, but, as we showed in §5.2, they rely on a relatively large[10] dataset of texts that were manually identified as CFM texts, and human-written texts to serve as 'negative' examples, and it is difficult to get the human-written samples right. It would take time, expertise, and resources to develop such a dataset. And we might want to detect a CFM before it publishes even 2K "news". The synthetic examples could also come from multiple CFMs, making the classification task even harder. It is telling that the classifier developed by OpenAI, presumably aiming to detect only OpenAI models, was soon shut down due to low accuracy (Kirchner et al., 2023).

Concerning watermarking, in the CFM case, it is safe to assume that the deployers of such models would likely try to remove or obfuscate the watermarking, which is relatively easy to do by altering the generation strategy. And the public easy-to-use tools highlighting the statistical 'evidence' of the watermark[11] could be used not only by those looking to detect CFMs, but also the CFM operators, to obscure the evidence.

We conclude that at the moment there are no practical options for detecting news 'content farms' in the wild, and both the open-source LLM community and providers of 'closed' LLMs need to consider ways to address that. We are hopeful that some combination of centralized watermarking effort and further development of detection methods could provide a working solution in the future.

Since the hardest problem is identifying the base LLM, one policy direction to consider would be to (a) mandate watermarking, ideally built-into-model-weights, as a pre-requisite for either commercial deployment of LLMs or their open-source publication, (b) maintain a public watermark detection library, allowing to identify a candidate LLM given a text sample. This would not be bullet-proof, but it would raise both the costs of avoiding detection and the awareness of the problem in the NLP community.

8 Conclusion

In this work we have shown how creating a 'content farm' generative model for news-like text can be easy, even though we started with a relatively old LLM and a language it was not originally meant for. After fine-tuning Llama 65B on only 40K Italian news texts, native speakers of Italian have only $\approx 64\%$ accuracy on the synthetic news text detection task.

We show that the current approaches to automatic text detection, based on token likelihood and supervised classification, outperform human raters in the synthetic text detection, but they would all be impractical in the real world, since they require access to the token likelihood information or a large training dataset. We further consider the proxy model approach, and we find that it works well even with little data for fine-tuning, but only if it is known which base LLM was used. Our study highlights the urgency of further work on developing model-agnostic methods of synthetic text detection.

[10] Verma et al. (2023) used 30k samples for English alone.

[11] E.g. demo by Kirchenbauer et al. (2023) at https://huggingface.co/spaces/tomg-group-umd/lm-watermarking

9 Limitations

Generalizability to other languages. We present a case study on a single language, and do not intend to claim that it is possible to generate plausible-sounding text in any language, by fine-tuning a mostly-English model like Llama. But our results suggest that it *may* be possible, at least for languages with a similar level of coverage in datasets used for training LLMs. More research is needed to establish both the factors impacting the success of such transfer, and better methods to detect synthetic texts.

For the original Llama, according to Touvron et al. (2023a, p.2), it was exposed to Italian Wikipedia in pre-training. Italian Wikipedia currently has about 500K articles.[12] For other sources included in Llama, such as C4 (Raffel et al., 2020), we cannot exclude the possibility that there was some Italian – but deliberate effort was made to filter out non-English texts, and so we assume that there was at least not much contamination. Other languages in Llama with about the same amount of Wikipedia data as Italian are Polish and Dutch ($\approx$ 500K articles). In the $\approx$ 400K range there are Spanish and Portuguese, and at about $\approx$ 300K – Russian and Swedish. A future study could explore how the amount of Wikipedia data, the amount of fine-tuning data, and the typological distance from English impact the success of the transfer.

It is of course also possible and likely that a CFM developer aiming for a specific language would start with an LLM that is multilingual by design, such as BLOOM (Scao et al., 2022) or AYA (Üstün et al., 2024), and probably get even better transfer.

Other methods of detecting synthetic text. In the scope of this paper, we experimented with two statistical detection methods (based on likelihood scores and supervised detection), but there are others, including more sofisticated likelihood based approaches (Su et al., 2023) that however share the same dependency on the likelihood of the original models and therefore limitations of those we experiment with. This does not invalidate our conclusions and the general answer to our research question, but it could be expanded in the future work.

Limitations of the human evaluation. Our selection of human raters was based solely on Italian as their native language. Future work could investigate whether the results would differ across different occupations and education levels, and with different kinds of synthetic and real articles.

Our human evaluation protocol considers the setting where the model is prompted with the first 30 tokens of a real human-written article, because the model is not trained using the articles headlines but just to generate news, to make the adaptation from English to Italian simpler. Another scenario to be tested in future work is when the model is prompted with headlines (authored by the content farm owner or auto-generated). That could affect the quality of the generated text or the ease of its detection.

Finally, our study focuses on the possibility to create plausible-sounding news-like text that could be used by "content farms", rather than text created for specific misinformation campaigns or to spread conspiracy theories. It is possible that, similarly to human-authored fake news, the human raters would be more likely to doubt the authenticity of the article when it had some big factual claims that were easy to check. This factor also remains to be explored in future work.

The best case scenario with a CFM is that the users would soon realize that the site is fake, and leave – but even in that case they would already have wasted time and resources, and potentially increased their digital footprint because of tracking on the CFM website. Further harms from misinformation, targeted ads etc. are beyond the scope of this study, and will require a more detailed investigation of various types of factually problematic content, deliberate attempts to introduce certain narratives, awareness and training of the users, etc.

Detection of API-based models. This study focused on the scenario where the 'content farm' used its own model, created by fine-tuning a publicly available high-performing base LLM. It could of course also use an external API service, which would make its task even easier technically.

Famously, GPT-2 (Radford et al., 2019) initially came with warnings about it being "too dangerous to release", precisely because of the danger of synthetic 'news' (Wakefield, 2019). Section 7.4 in the GPT-3 (Brown et al., 2020) report is dedicated to synthetic news generation and the finding that humans detect such 200-word GPT-3 texts in the 52%-76% accuracy range. At that time, too, there was coverage of the dangers of synthetic news (e.g. Mak, 2019; Knight, 2021), but no policies ensued.

[12] https://en.wikipedia.org/wiki/Wikipedia:Multilingual_statistics

Now that OpenAI offers the most popular generative AI services, its proposed solution is its Terms of Service, presumably enforced via constant monitoring of the API use by all users. Its current Terms of Service broadly prohibit "any harmful, illegal, or abusive activity", but it is not immediately clear which definitions for 'harm' and 'legality' must be followed, and whether the news 'content farms' websites are covered. The most directly relevant clause currently[13] seems to be representing "that Output was human-generated when it was not". We do not know how this is enforced, and how many bad actors are successfully stopped with API-level controls – but clearly not all of them.[14]

10 Broader Impacts

Impact on society. This work aims to highlight a potential problem for the information infrastructure of worldwide communities that may consider themselves safe from plausible-looking synthetic text due to the lack of high-quality monolingual models for their languages. We show that a relatively old Llama model, exposed only to Italian Wikipedia and 40K news articles for fine-tuning, is sufficient for generating very plausible-looking synthetic news in Italian, and there are no practical solutions for detecting such text. We hope that this work would spark similar investigations for other languages, and highlight the urgency of development of reliable and model-agnostic methods for detecting synthetic text.

In particular, we hope to draw the attention to the fact that at present, the most authoritative source of information about the extent of the problem with the 'content farm' news websites seems to be the aforementioned NewsGuard reports (Sadeghi and Arvanitis, 2023; Sadeghi et al., 2024). They are based on extensive expert research and manually vetting different news outlets. However, NewsGuard is a paid service, and, ideally, the society would not make something like news a privilege of the wealthier citizens.[15] The problem of the current huge wave of AI-generated 'news' originates in the field of NLP, and our field is also best positioned to work on the problems of detecting synthetic texts and assisting with their reporting.

Human and computational resources. This work is based on the publicly available models (Radford et al., 2019; Touvron et al., 2023a,b; Raffel et al., 2020; Sarti and Nissim, 2022; Jiang et al., 2023) and resources (Narayan et al., 2018; Mattei et al., 2020), and documents its emissions (App. H), annotation procedure and compensation to the human raters (§4). The code to reproduce our experiments accompanies the submission will be publicly available with the publication of the paper.

References

Giovanni Bonisoli, Maria Pia di Buono, Laura Po, and Federica Rollo. 2023. Dice: A dataset of italian crime event news. In *SIGIR '23: The 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, Taipei, Taiwan, July 23 - 27, 2023*. ACM.

Jack Brewster, Macrina Wang, and Coalter Palmer. 2023. Plagiarism-Bot? How Low-Quality Websites Are Using AI to Deceptively Rewrite Content from Mainstream News Outlets - Misinformation Monitor: August 2023.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Souradip Chakraborty, Amrit Singh Bedi, Sicheng Zhu, Bang An, Dinesh Manocha, and Furong Huang. 2023. On the possibilities of AI-generated text detection.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco

[13] https://openai.com/policies/terms-of-use, accessed on Feb 10 2023.

[14] For example, Gizmodo identified a fake story about the death of Joe Biden that started with "*I'm sorry, I cannot complete this prompt as it goes against OpenAI's use case policy on generating misleading content. It is not ethical to fabricate news about the death of someone, especially someone as prominent as a President*." (DeGeurin, 2023)

[15] Admittedly, full-length high-quality news have always been a paid service by professional outlets, but there are at least two new factors that could make the problem worse nowadays: (a) the news aggregators have long habituated the public to only reading headlines and snippets, which at least partly substitute reading the full articles (Dellarocas et al., 2016), (b) people now may only read the snippets even before sharing them on social media (Holmström et al., 2020). Since the content farms are proliferating (§1), those snippets must have a click rate high enough to be profitable. To ensure that, a lot more people must be exposed to the junk snippets, and hence may absorb at least some misinformation from them.

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*, pages 8440–8451, Online. Association for Computational Linguistics.

Mack DeGeurin. 2023. No, Biden Isn't Dead: AI Content Farms Are Pumping Out Fake Stories.

Chrysanthos Dellarocas, Juliana Sutanto, Mihai Calin, and Elia Palme. 2016. Attention Allocation in Information-Rich Environments: The Case of News Aggregators. *Management Science*, 62(9):2543–2562.

Heather Desaire, Aleesa E. Chua, Min-Gyu Kim, and David Hua. 2023. Accurately detecting AI text when ChatGPT is told to write like a chemist. *Cell Reports Physical Science*, 4(11):101672.

Mahdi Dhaini, Wessel Poelman, and Ege Erdogan. 2023. Detecting ChatGPT: A Survey of the State of Detecting ChatGPT-Generated Text. In *Proceedings of the 8th Student Research Workshop Associated with the International Conference Recent Advances in Natural Language Processing*, pages 1–12, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Pierre Fernandez, Antoine Chaffin, Karim Tit, Vivien Chappelier, and Teddy Furon. 2023. Three Bricks to Consolidate Watermarks for Large Language Models. In *2023 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–6.

J.L. Fleiss et al. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Soumya Suvra Ghosal, Souradip Chakraborty, Jonas Geiping, Furong Huang, Dinesh Manocha, and Amrit Bedi. 2023. A survey on the possibilities & impossibilities of AI-generated text detection. *Transactions on Machine Learning Research*. Survey Certification.

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023. How Close is ChatGPT to Human Experts? Comparison Corpus, Evaluation, and Detection.

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.

Jesper Holmström, Daniel Jonsson, Filip Polbratt, Olav Nilsson, Linnea Lundström, Sebastian Ragnarsson, Anton Forsberg, Karl Andersson, and Niklas Carlsson. 2020. Do we read what we share? analyzing the click dynamic of news articles shared on Twitter. In *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '19, pages 420–425, New York, NY, USA. Association for Computing Machinery.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Ganesh Jawahar, Muhammad Abdul-Mageed, and Laks V. S. Lakshmanan. 2020. Automatic detection of machine generated text: A critical survey.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. 2023. A watermark for large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 17061–17084. PMLR.

Jan Hendrik Kirchner, Lama Ahmad, Scott Aaronson, and Jan Leike. 2023. New AI classifier for indicating AI-written text.

Will Knight. 2021. AI Can Write Disinformation Now—and Dupe Human Readers. *Wired*.

Rohith Kuditipudi, John Thickstun, Tatsunori Hashimoto, and Percy Liang. 2023. Robust Distortion-free Watermarks for Language Models.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Peixuan Li, Pengzhou Cheng, Fangqi Li, Wei Du, Haodong Zhao, and Gongshen Liu. 2023. PLMmark: A Secure and Robust Black-Box Watermarking Framework for Pre-trained Language Models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(12):14991–14999.

Wenxiong Liao, Zhengliang Liu, Haixing Dai, Shaochen Xu, Zihao Wu, Yiyang Zhang, Xiaoke Huang, Dajiang Zhu, Hongmin Cai, Quanzheng Li, Tianming Liu, and Xiang Li. 2023. Differentiating ChatGPT-Generated and Human-Written Medical Texts: Quantitative Study. *JMIR Medical Education*, 9:e48904.

Yikang Liu, Ziyin Zhang, Wanyang Zhang, Shisen Yue, Xiaojing Zhao, Xinyuan Cheng, Yiwen Zhang, and Hai Hu. 2023. ArguGPT: Evaluating, understanding and identifying argumentative essays generated by GPT models.

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.

Sasha Luccioni, Victor Schmidt, Alexandre Lacoste, and Thomas Dandres. 2019. Quantifying the carbon emissions of machine learning. In *NeurIPS 2019 Workshop on Tackling Climate Change with Machine Learning*.

Aaron Mak. 2019. When Is Technology Too Dangerous to Release to the Public? *Slate*.

Lorenzo De Mattei, Michele Cafagna, Felice Dell'Orletta, Malvina Nissim, and Albert Gatt. 2020. CHANGE-IT @ EVALITA 2020: Change headlines, adapt news, generate (short paper). In *Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020), Online event, December 17th, 2020*, volume 2765 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. 2023. DetectGPT: Zero-shot machine-generated text detection using probability curvature.

Sandra Mitrović, Davide Andreoletti, and Omran Ayoub. 2023. ChatGPT or Human? Detect and Explain. Explaining Decisions of Machine Learning Model for Detecting Short ChatGPT-generated Text.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

OpenAI. 2022. Introducing ChatGPT.

OpenAI. 2023. GPT-4 Technical Report.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *Technical report*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1).

Tate Ryan-Mosley. 2023. Junk websites filled with AI-generated text are pulling in money from programmatic ads. *MIT Technology Review*.

Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. 2023. Can AI-generated text be reliably detected?

McKenzie Sadeghi and Lorenzo Arvanitis. 2023. Rise of the Newsbots: AI-Generated News Websites Proliferating Online.

McKenzie Sadeghi, Lorenzo Arvanitis, Virginia Padovese, Giulia Pozzy, Sara Badilini, Chiara Vercellone, Madeline Roache, Macrina Wang, Jack Brewster, Natalie Huet, Becca Schimmel, Andie Slomka, Leonie Pfaller, Louise Vallee, and Natalie Adams. 2024. Tracking AI-enabled Misinformation: Over 650 'Unreliable AI-Generated News' Websites (and Counting), Plus the Top False Narratives Generated by Artificial Intelligence Tools.

Siddharth Samsi, Dan Zhao, Joseph McDonald, Baolin Li, Adam Michaleas, Michael Jones, William Bergeron, Jeremy Kepner, Devesh Tiwari, and Vijay Gadepally. 2023. From words to watts: Benchmarking the energy costs of large language model inference.

Gabriele Sarti and Malvina Nissim. 2022. It5: Large-scale text-to-text pretraining for italian language understanding and generation.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, Dragomir Radev, Eduardo González Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady Elsahar, Hamza Benyamina, Hieu Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Frohberg, Joseph Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro Von Werra, Leon Weber, Long Phan, Loubna Ben allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, María Grandury, Mario Šaško, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A. Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto Luis López, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, Shayne Longpre, Somaieh Nikpoor, Stanislav Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin

Heinzerling, Chenglei Si, Davut Emre Taşar, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M. Saiful Bari, Maged S. Al-shaibani, Matteo Manica, Nihal Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Fevry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiangru Tang, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Yallow Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre François Lavallée, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aurélie Névéol, Charles Lovering, Dan Garrette, Deepak Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Jordan Clive, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shachar Mirkin, Shani Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdeněk Kasner, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ana Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Ajibade, Bharat Saxena, Carlos Muñoz Ferrandis, Danish Contractor, David Lansky, Davis David, Douwe Kiela, Duong A. Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatima Mirza, Frankline Ononiwu, Habib Rezanejad, Hessie Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jesse Passmore, Josh Seltzer, Julio Bonis Sanz, Livia Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, Muhammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nour Fahmy, Olanrewaju Samuel, Ran An, Rasmus Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas Wang, Sourav Roy, Sylvain Viguier, Thanh Le, Tobi Oyebade, Trieu Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Singh, Benjamin Beilharz, Bo Wang, Caio Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel León Periñán, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Imane Bello, Ishani Dash, Jihyun Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthik Rangasai Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, Maria A. Castillo, Marianna Nezhurina, Mario Sänger, Matthias Samwald, Michael Cullan, Michael Weinberg, Michiel De Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patrick Haller, Ramya Chandrasekhar, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S. Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aroonsiri, Srishti Kumar, Stefan Schweter, Sushil Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yash Shailesh Bajaj, Yash Venkatraman, Yifan Xu, Yingxin Xu, Yu Xu, Zhe Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. 2022. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model.

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.

Yuki Takezawa, Ryoma Sato, Han Bao, Kenta Niwa, and Makoto Yamada. 2023. Necessary and Sufficient Watermark for Large Language Models.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas

Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models.

Chris van der Lee, Albert Gatt, Emiel van Miltenburg, and Emiel Krahmer. 2021. Human evaluation of automatically generated text: Current trends and best practice guidelines. *Computer Speech & Language*, 67:101151.

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.

Vivek Verma, Eve Fleisig, Nicholas Tomlin, and Dan Klein. 2023. Ghostbuster: Detecting text ghostwritten by large language models.

Jane Wakefield. 2019. 'Dangerous' AI offers to write fake news. *BBC News*.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Chenxi Whitehouse, Osama Mohammed Afzal, Tarek Mahmoud, Alham Fikri Aji, and Preslav Nakov. 2023. M4: Multi-generator, multi-domain, and multi-lingual black-box machine-generated text detection.

Yihan Wu, Zhengmian Hu, Hongyang Zhang, and Heng Huang. 2023. DiPmark: A Stealthy, Efficient and Resilient Watermark for Large Language Models.

KiYoon Yoo, Wonhyuk Ahn, and Nojun Kwak. 2023. Advancing Beyond Identification: Multi-bit Watermark for Large Language Models.

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model. *arXiv preprint arXiv:2402.07827*.

pretrained	CHANGE-it fine-tuning samples				HuggingFace model
	3,981	7,862	40,000	127,392	
llama-1-7b	-	-	llama-7b_it	-	huggyllama/llama-7b
llama-1-65b	-	-	llamam-65b_it	-	huggyllama/llama-65b
llama-2-7b	*llama-2-7b_it_3981*	*llama-2-7b_it_7862*	-	*llama-2-7b_it*	meta-llama/Llama-2-7b
llama-2-13b	*llama-2-13b_it_3981*	*llama-2-13b_it_7862*	-	*llama-2-13b_it*	meta-llama/Llama-2-13b
mistral	*mistral_it_3981*	*mistral_it_7862*		*mistral_it*	meta-llama/Llama-2-70b

Table 4: Model naming based on pretrained model and number of fine-tuning samples. The fine-tuning samples number refers to the training set of CHANGE-it (Mattei et al., 2020) dataset. HuggingFace model names correspond to the current links on HuggingFace hub (e.g. for *meta-llama/Llama-2-7b* the pre-trained model comes from `https://huggingface.co/meta-llama/Llama-2-7b`).

A Model Sources and Naming Scheme

All LLMs used in this study are listed in Table 4. We also add name of the models in Huggingface-Hub, e.g. for *meta-llama/Llama-2-7b* the pre-trained model comes from `https://huggingface.co/meta-llama/Llama-2-7b`.

B Fine-tuning Details

B.1 Llama 1

We fine-tune both Llama 7b and 65b models on a randomly chosen 40K subset of the CHANGE-it news dataset. The articles are arranged in training sequences composed of 128 tokens, adding subsequent segments one after the other.

The training was performed on 8 nodes, each with 4 V100 GPUs with 16GB VRAM. The effective batch size is 128, real batch size 2, 64 accumulation steps. The maximum learning rate is 0.0005, with a one-cycle scheduler without warmup.

Due to memory constraints, instead of the most effective AdamW optimizer we use simple Stochastic Gradient Descent. We train for 60,000 steps (120,000 samples split into 3 epochs of 40,000) saving checkpoints every 10,000 steps (20,000 samples). This takes approximately 3 days.

B.2 Llama 2 and Mistral

We fine-tune Llama 2 7b and 13b as well as Mistral models on the full training set of the CHANGE-it news dataset. With a length up to 2048 tokens.

The training is performed on 4 nodes, each with 4 A100 GPUs with 64GB VRAM. The effective batch size is 256, real batch size 4 with 4 accumulation steps. The maximum learning rate is 1.e-5, with a one-cycle scheduler without warm-up.

We use AdamW optimizer with FSDP, we train for three epochs in bfloat16 to limit the memory needed and we perform full fine-tuning without using any parameter efficient technique.

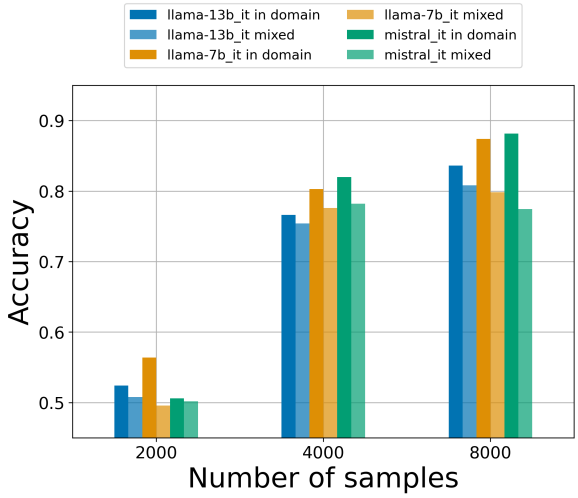


Figure 5: Accuracy of classifier based on RoBERTa-large for human/synthetic text classification task, for synthetic texts generated by three LLMs fine-tuned on CHANGE-it. The classifier was trained on 50% synthetic texts and either 50% CHANGE-it texts (*in domain*), or 25% texts from CHANGE-it and 25% from DICE (*mixed source*). Classification is only successful at at least 4K labeled samples, and the *mixed source* scenario is consistently more challenging.

For the fine-tuning on subsets of the CHANGE-it dataset we keep most things equal to the longer training-set.

For the XSUM dataset the same setting is kept almost identical.

B.3 Fine Tuning Supervised Synthetic Text Detectors

We fine-tune two classifiers to identify real or synthetic texts, RoBERTa-large and XLM-RoBERTa-Large. The models are trained on three different dataset size, 2k, 4k, 8k and two data mixing.

The max learning rate is 5.e-5 we use the a batch

generator	n samples	roberta-large	xlm-large
In Domain			
llama-2-13b_it	2000	0.52	0.73
	4000	0.77	0.81
	8000	0.84	0.90
llama-2-7b_it	2000	0.56	0.59
	4000	0.80	0.86
	8000	0.87	0.92
mistral_it	2000	0.51	0.52
	4000	0.82	0.81
	8000	0.88	0.84
Mixed Source			
llama-2-7b_it	2000	0.51	0.53
	4000	0.75	0.79
	8000	0.81	0.88
llama-2-13b_it	2000	0.50	0.61
	4000	0.78	0.86
	8000	0.80	0.90
mistral_it	2000	0.50	0.55
	4000	0.78	0.82
	8000	0.77	0.83

Table 5: Accuracy achieved by supervised classification models when trained to classify real and synthetic text generated by the various fine-tuned models. We report the accuracy on *in domain* (where all synthetic texts are generated from the same source) and on *mixed source* data (where half of the human written texts come from a different Italian news outlet). Our results suggest that supervised classification of synthetic texts critically depends on the availability of a large (at least 4K in this case) labelled training set, particularly in the *mixed source* scenario.

size of 128 and a linear decaying learning rate without warmup. The rest of the parameters are the daults parameters used by Huggingface Trainer.

The performance is reported in Figure 4 for XLM-RoBERTa and Figure 5 for RoBERTa.

The detailed numerical results are also listed in Table 5.

C Alternative Threshold for Human Evaluation Metric

To compute the readers accuracy on identifying machine generated texts, we threshold the average score assigned to a sample to obtain a binary label. In §4 we show the results using 3 as a threshold, which is the mean possible rating, but we find that the results are similar when the threshold is the **scaled mean**: the mean score of all questions in a 100-question survey. Table 6 shows all results with this threshold, and they are well aligned with Table 1.

model	accuracy	std
Llama 7B pretrain	85.3	6.1
Llama 7B finetune	73.7	8.8
Llama 65B pretrain	72.6	3.9
Llama 65B finetune	65.3	16.9

Table 6: Accuracy achieved by human raters in assessing human-written versus machine generated news. We report the overall accuracy and the standard deviation.

D Generation examples

This appendix lists further examples of LLMs pretrained mostly on English switching from English to Italian when generating the text, as well as the same examples where our CFMs don't do so. We show this for Llama 7b not-pretrained Table 8, Llama 7b fine-tuned Table 9, Llama 65b pre-trained Table 10 and Llama 65b pre-trained Table 11. The captions also report a break down of the output of the qualitative study we carried out, listing the amount and type of mistakes we spot.

Furthermore, Table 12 shows examples of the changes done by it5 when generating the modifications necessary to compute the *DetectGPT* score. These modifications can sometimes lchange meaning in Italian, although when the modified spans are only 1 or 2 words long they keep their original meaning.

Models Switching to English
Original Text: Due sbarchi in sole ventiquattr'ore a Lampedusa. Sono 84 gli immigrati giunti nell'isola tra venerdì e sabato e ospitati all'Hotspot di contrada Imbriacola. Tra loro ci sono anche I tre unici superstiti dell'ultimo naufragio avvenuto a 50 chilometri a nord est dalla Libia, costato la vita a 117 persone, che erano partite giovedì a bordo di un gommone sgonfiatosi a circa undici ore di navigazione. I passeggeri sono morti annegati, mentre I tre naufraghi sono riusciti a trovare un appiglio fino all'arrivo dei soccorsi italiani. I tre stranieri...
Pre-trained Llama: Due sbarchi in sole ventiquattr'ore a Lampedusa. Sono 84 gli immigrati giunti nell'isola tra venerdì e sabato e ospitati all'Hotspot di contrada Imbriacola. Tra loro ci sono anche I tre naufraghi di Tortuga, recuperati da un mercantile libanese. Non è ancora stata resa nota l'identità nazionale dei naufraghi, ma non mancano inchieste di persone che si sono rivolte ai carabinieri per cercare i loro cari. Lampedusa, 38 migrants - mostly from Tunisia - arrived safely on the island ...
Fine tuned epoch 6 Llama: Due sbarchi in sole ventiquattr'ore a Lampedusa. Sono 84 gli immigrati giunti nell'isola tra venerdì e sabato e ospitati all'Hotspot di contrada Imbriacola. Tra loro ci sono anche I tre bambini rimasti uccisi durante lo stesso sbarcho, due di questi non appartenevano a una famiglia, ma erano stati imbarcati separati dai genitori per il viaggio da Libia verso l'Italia. È stato appunto il maggiore Damiano Maiorino, di stanza a Lampedusa, a scoprire i morti. "Non sono morti cadendo in mare da una barca troppo affollata. Hanno subito commentato ...
Field
Original Text: L' Onu è pronta a fare la sua parte alla conclusione del vertice europeo, ma il segretariato non ha in agenda alcuna azione militare, anche se ci possono essere alcune similarità con l'operazione in Somalia ". In una intervista a Qn il vicesegretario generale dell'Onu Jan Eliasson apre "all'uso della forza" per fermare gli sbarchi dei clandestini ed evitare nuove stragi nel Mediterraneo. Ma, al tempo stesso, sottolinea "una differenza qualitativa" con l'operazione in Somalia: "L'attività dei pirati era solo criminale, mentre in Libia abbiamo un urgentissimo problema umanitario prima che criminale". Parlando dell'affondamento preventivo dei barconi nei porti libici...
Pre-trained Llama: L' Onu è pronta a fare la sua parte alla conclusione del vertice europeo, ma il segretariato non ha in agenda alcuna azione militare, anche se ci possono essere alcune iniziative sul campo degli aiuti umanitari", ha affermato un portavoce dell'organizzazione internazionale. Per il segretario generale, "l' Italia deve avere una posizione di leader ed essere la prima a proporre una risposta alla crisi europea, visto anche che l' Italia ha avuto un ruolo molto importante nei negoziati sul tema del contingente europeo". UNITED NATIONS, January 31 – Without waiting for the European Union's decision on...
Fine tuned epoch 6 Llama: L' Onu è pronta a fare la sua parte alla conclusione del vertice europeo, ma il segretariato non ha in agenda alcuna azione militare, anche se ci possono essere alcune azioni umanitarie" . Così parla il portavoce dell'organizzazione internazionale, Stephane Dujarric, che aggiunge: "Un vertice europeo è una questione interna". La conferma che l'Onu non è pronta a seguito di una intervista rilasciata dal segretario generale aggiunto, Jeffrey Feltman, a radio 24. "Non abbiamo notizie di alcuna iniziativa che possa condurre all'attuazione di un mandato militare contro l'islamic State sul territorio siriano, anche se ritengo che si stia...

Table 7: Real and generated sentences in Italian from Llama 65B pretrained and Llama 65B fine-tuned on Italian all starting from the same prompt. The sentences generated by Llama 65B pretrained are not detected by either DetectGPT or log-likelihood, despite being partially written in English – which could not happen in a real Italian news article.

Llama 7B Pre-Trained
Generated Text: I dodici ragazzi rimasti intrappolati con il loro allenatore nella grotta di Tham Luang, nella provincia di Chiang Rai, in Thailandia, per oltre due settimane, hanno lasciato oggi l'ospedale dove erano ricoverati. La polizia ha riferito che i ragazzi sono stati sottoposti a un test del DNA che ha dimostrato che tutti erano discendenti di due genitori, ma non vi è nessuna prova che uno o più ragazzi non sia rimasto tra loro durante la fuga dalla grotta. Il...
Translation: The 12 boys who were trapped with their coach in the Tham Luang cave in Thailand's Chiang Rai province for more than two weeks left the hospital today where they were hospitalized. Police reported that the boys underwent a DNA test that showed that all were descendants of two parents, but there is no evidence that one or more of the boys did not remain among them while escaping the cave. The
Issue type: Grammatical error.
Original Text: Se dovesse darsi un voto, lei si darebbe un "sette e mezzo". Sette romani su dieci, però, secondo un sondaggio di Repubblica, non le darebbero nemmeno la sufficienza. Ad un anno e mezzo dal suo insediamento, è la dimostrazione più esplicita del limpido disaccordo che la popolazione ha con la sua amministrazione. From the review of A. Kersalme's "Odysseus" on the site of French journal "La Revue de Belles Lettres" (98th issue, March 2019). From the review of M....
Translation: If she had to give herself a grade, she would give herself a "seven and a half." Seven out of ten Romans, however, would not even give her a sufficiency, according to a Repubblica poll. A year and a half after taking office, it is the most explicit demonstration of the limpid disagreement the population has with her administration. From the review of A. Kersalme's "Odysseus" on the site of French journal "La Revue de Belles Lettres" (98th issue, March 2019). From the review of M....
Issue type: Factual error.
Generated Text: Passerà in carcere da un minimo di 42 anni fino all'ergastolo. Si è chiuso con la lettura di una sentenza durissima da parte del giudice della Superior Court di Los Angeles. La condanna dura fino al 2019, ma la rischia se si insulta o rinnega ancora una volta le sue azioni. Il procuratore generale della California, Kamala Harris, ha anche chiesto per lui la revoca di libertà condizionata. Il giovane è stato arrestato a Hong Kong nel maggio 2014....
Translation: He will spend from a minimum of 42 years to life in prison. It ended with the reading of a very harsh sentence by the Los Angeles Superior Court judge. The sentence lasts until 2019, but he risks it if he insults or disavows his actions again. California Attorney General Kamala Harris also sought revocation of parole for him. The young man was arrested in Hong Kong in May 2014...
Issue type: Grammatically correct, but does not sound natural in Italian and shows factual errors.

Table 8: Examples of mistakes made by pretrained Llama 7B. The smaller Llama 7b without fine-tuning shows several kinds of weaknesses generating both factual to grammatical errors. Out of 25 samples manually examined for this model, we find 0 contradicting the prompt, 5 switching to English, 8 unnatural Italian expressions and 6 grammatical errors.

Llama 7B Fine-Tuned
Generated Text: Passa per sette punti il programma di Matteo Salvini per bloccare I flussi e dare maggiore sicurezza al Paese. Il blocco dei porti a navi che battono bandiera di uno stato membro dell'Unione europea e non il Regno Unito, come annunciato da Salvini, sarebbe contro il diritto e il trattato, secondo l'ex ministro della Giustizia Franco Frattini. "Vi è un'interpretazione di diritto internazionale. A me pare che sia totalmente sbagliata", ha detto il politico liberale di Palazzo Chigi. "Sotto...
Translation: It passes for seven points in Matteo Salvini's program to stop the flows and give more security to the country. Blocking ports to ships flying the flag of a European Union member state and not the United Kingdom, as announced by Salvini, would be against law and treaty, according to former Justice Minister Franco Frattini. "There is an interpretation of international law. It seems to me that it is totally wrong," said the liberal politician from Palazzo Chigi. "Under...
Issue type: Factual errors.
Original Text: Alla vigilia dell'incontro, a Bruxelles, tra Jean-Claude Juncker e il presidente del Consiglio Giuseppe Conte, che sarà accompagnato dal ministro dell'Economia Tria, il commissario Ue Pierre Moscovici usa tona concilianti per suggerire che l'Italia non dovrebbe fare la spola tra l'Europa e le banche a credito. "Sarà difficile", dice il francese, "se la Banca centrale germanica non pone il suo veto alla pesca in una sola mano per un paese come l'Italia che sta in cerca di finanziamenti". "La...
Translation: On the eve of a meeting in Brussels between Jean-Claude Juncker and Prime Minister Giuseppe Conte, who will be accompanied by Economy Minister Tria, EU Commissioner Pierre Moscovici uses conciliatory tones to suggest that Italy should not shuttle between Europe and banks on credit. "It will be difficult," says the Frenchman, "if the German Central Bank does not veto single-handed fishing for a country like Italy that is looking for financing." "The...
Issue type: Unnatural expression.
Generated Text: Forti temporali si sono abbattuti nella mattinata di oggi su Milano e sul nord-ovest della Lombardia, spazzando via il caldo degli ultimi giorni. Il nubifragio ha creato qualche rallentamento sulle strade. Poche colture e alcune piantine. Questi sono i principali danni dei fenomeni atmosferici che hanno colpito l'Italia. L'Italia centro-meridionale ha visto 304.363 colture danneggiate, quasi 15mila piantine e 1.532 migliaia di alberi, una pioggia netta di oltre 10mila mila chili sul versante del Po e una coltura danneggiata...
Translation: Strong thunderstorms hit Milan and northwestern Lombardy in the morning today, sweeping away the heat of the past few days. The cloudburst created some slowdowns on roads. Few crops and a few seedlings. These are the main damages of the weather phenomena that hit Italy. South-central Italy saw 304,363 crops damaged, nearly 15,000 seedlings and 1,532,000 trees, a net rainfall of more than 10,000,000 kilograms on the Po River side, and a damaged crop...
Issue type: Factual error.

Table 9: Examples of mistakes made by Llama 7B fine-tuned on Italian. The smaller Llama 7b after fine-tuning generates flowing text but the facts are anyway less accurate than for larger models. Out of 25 samples manually examined for this model, we find 5 contradicting the prompt, 0 switching to English, 5 unnatural Italian expressions and 6 grammatical errors.

Llama 65B Pre-Trained
Generated Text: In centro Italia continuano le scosse e il numero dei morti sale. Paesi afoni e distrutti, palazzi accartocciati, mozziconi di chiese, rovine sinistre che fissano le piazze. E I corpi stanno fuori dalle case, senza casa, senza neanche la tomba che li ricuopre. Le indagini del Giornalista Enrico Lucci e le testimonianze dei superstiti. The BBC's Gavin Lee reports from Norcia, the central Italian town hardest hit by the quake. Italian Prime Minister Matteo Renzi has visited one of...
Translation: In central Italy, the tremors continue and the death toll rises. Aphonious and destroyed villages, crumpled buildings, church butts, sinister ruins staring into squares. And The bodies stand outside homes, homeless, without even the grave to cover them. Journalist Enrico Lucci's investigation and survivor testimony. The BBC's Gavin Lee reports from Norcia, the central Italian town hardest hit by the quake. Italian Prime Minister Matteo Renzi has visited one of...
Issue type: Grammatically correct, but does not sound natural in Italian.
Original Text: Due sbarchi in sole ventiquattr'ore a Lampedusa. Sono 84 gli immigrati giunti nell'isola tra venerdì e sabato e ospitati all'Hotspot di contrada Imbriacola. Tra loro ci sono anche I tre naufraghi di Tortuga, recuperati da un mercantile libanese. Non è ancora stata resa nota l'identità nazionale dei naufraghi, ma non mancano inchieste di persone che si sono rivolte ai carabinieri per cercare I loro cari. Lampedusa, 38 migrants - mostly from Tunisia - arrived safely on the island in...
Translation: Two landings in just twenty-four hours on Lampedusa. A total of 84 immigrants arrived on the island between Friday and Saturday and were housed at the contrada Imbriacola Hotspot. Among them are also The three shipwrecked Tortuga, recovered by a Lebanese merchant ship. The national identity of the castaways has not yet been released, but there is no shortage of inquiries from people who have turned to the Carabinieri to search for their loved ones. Lampedusa, 38 migrants - mostly from Tunisia - arrived safely on the island in...
Issue type: Switch to English
Generated Text: I numeri fanno spavento. Cinquemila disperati sbarcati solo nelle ultime quarantott'ore, quasi 65mila nei primi sei mesi dell'anno. Che l' operazione "Mare Nostrum", varata dopo la strage di Lampedusa, fosse riuscita a fermare il flusso di barche di migranti, lo abbiamo rilevato già da pochi mesi. Che la spinta verso l'Italia del prossimo potesse crescere di conseguenza, lo sapevamo. Che si dovesse progettare una strategia efficace e responsabile, in grado di dire dove e come, giustamente, stava l' opportunità...
Translation: The numbers are frightening. Five thousand desperate people landed in the last forty-eight hours alone, nearly 65,000 in the first six months of the year. That operation "Mare Nostrum," launched after the Lampedusa massacre, had succeeded in stopping the flow of migrant boats, we noted a few months ago. That the push to Italy of the next could grow as a result, we knew. That an effective and responsible strategy had to be designed that could tell where and how, rightly, the 'opportunity lay....
Issue type: Contradicts the prompt.

Table 10: Examples of subtle mistakes made by pretrained Llama 65B. While the quality of the language is almost on par with the prompts, written by professional journalists (for these short examples), the content is largely made up although in a realistic way. Out of 25 samples manually examined for this model, we find 2 contradicting the prompt, 3 switching to English, 7 unnatural Italian expressions and 2 grammatical errors.

Llama 65B Fine-Tuned
Generated Text: Gabriele Gravina è stato eletto presidente della Figc. Ha ottenuto il 92,2% dei voti (al primo turno serviva il 75%). Si chiude la gestione commissariale di Roberto Fabbricini, durata 9 mesi. L'ultimo presidente eletto era Carlo Tavecchio nel 2014, che aveva ottenuto il 63%. Gravina: "Riconoscendomi orgogliosamente l'apostrofe di figlio della Sicilia e della Normandia che le mie origini e la mia vita mi fanno, ho iniziato la mia esperienza nel calcio. Un calcio che ho giocato da bambino,...
Translation: Gabriele Gravina has been elected president of Figc. He has obtained 92.2% of the votes (on first round he needed 76%), This ends the Roberto Fabbricini's technical administration, which lasted 9 months. The last president to be elected was Carlo Tavecchio in 2014, he scored 63%. Gravina: "Proudly acknowledging the apostrophe of son of Sicily and Normandy that my origins and my life make me, I began my experience in soccer. A soccer I played as a child,..."
Issue type: Factual errors.
Original Text: "È superficiale dire che è risorta la Dc ". Parola di Francesco Rutelli che, intervistato dal Corriere, ripercorre le tappe della vita della Margherita, partito in cui hanno militato sia lui che Bertinotti. Rutelli l'ha lasciato da qualche anno e adesso è presidente dei Liberali per l'Italia, la lista che recentemente ha lanciato al Senato. E Bertinotti è tornato alla Dc per partecipare alle primarie del partito di D'Alema e Bersani, l'unico che ha il voto dei sostenitori della...
Translation: "It is superficial to say that the DC has risen again." Word of Francesco Rutelli, who, interviewed by Corriere, traces the stages of the life of Margherita, a party in which both he and Bertinotti militated. Rutelli left it a few years ago and is now president of Liberals for Italy, the list he recently launched in the Senate. And Bertinotti returned to the DC to participate in the primaries of D'Alema and Bersani's party, the only one that has the supporter vote of the...
Issue type: Factual errors.
Generated Text: Dopo 19 giorni, grazie all'accordo trovato dall'Europa, I 49 migranti di Sea Watch e Sea Eye sono sbarcati nel porto maltese di La Valetta, dove sono stati trasportati a bordo di autobus. Il ministro Migranti e Strade del Popolo Gabriele Toccafondi nei giorni scorsi ha incontrato a Parigi il ministro degli Esteri francese Jean-Yves Le Drian, il ministro dell'interno Christophe Castaner, il ministro della giustizia Nicole Belloubet e il ministro dell'istruzione Jean-Michel Blanquer. Si tratta di una missione di...
Translation: After 19 days, thanks to the agreement found by Europe, The 49 migrants from Sea Watch and Sea Eye landed in the Maltese port of La Valetta, where they were transported aboard buses. Migrants and People's Roads Minister Gabriele Toccafondi in recent days met in Paris with French Foreign Minister Jean-Yves Le Drian, Interior Minister Christophe Castaner, Justice Minister Nicole Belloubet, and Education Minister Jean-Michel Blanquer. This is a mission of...
Issue type: Factual errors.

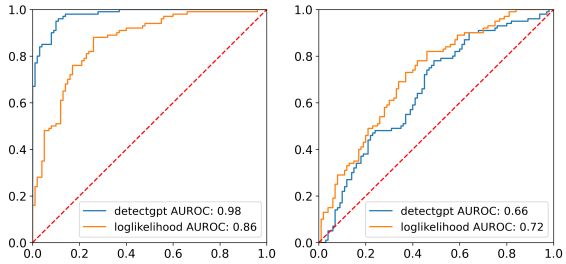
Table 11: Examples of subtle mistakes made by Llama 65B fine-tuned on Italian. While the quality of the language is almost on par with the prompts, written by professional journalists (for these short examples), the content is largely made up although in a realistic way. Out of 25 samples manually examined for this model, we find 0 contradicting the prompt, 0 switching to English, 1 unnatural Italian expressions and 4 grammatical errors.

Field	Content
Original Text:	Dato il seguente titolo di un articolo di giornale scrivi l'articolo. ### Titolo:"Nel Pd niente conte interne". Ma su un sito i nomi dei pro M5S ### Articolo: Alla vigilia della direzione che potrebbe, ma non necessariamente, sancire la separazione con gli scissionisti guidati da Renzi, il Pd cerca di chiudere la partita che tra poco si apre. La maggioranza del partito, insistono al Nazareno, per ora "non intende partecipare ad un conto interne",
Translation	Given the following title of a newspaper article, write the article. ### Title:"No internal counts in the Democratic Party". But on a site the names of the M5S pros ### Article: On the eve of the direction that could, but not necessarily, sanction the separation with the splinterists led by Renzi, the Democratic Party tries to close the game that will soon open . The majority of the party, the Nazareno insists, for now "does not intend to participate in an internal account"
Modification 1:	Dato il seguente titolo di anteprima di giornale scrivi lárticolo.### Titolo: "Nel Pd niente esodati, Ma su un sito I nomi dei pro M5S ### Articolo: Alla vigilia della direzione che potrebbe, ma non necessariamente, sancire la rottura definitiva con gli scissionisti guidati da Renzi, il pd cerca il suo partner di governo dietro la partita m5s che da poco si apre. La maggioranza del partito, insistono al Nazareno, per ora "non intende partecipare ad un conto interne"
Translation:	Given the following newspaper preview title, write an article.### Title: "No exoduses in the Democratic Party, but on a site the names of the M5S pros ### Article: On the eve of the direction that could , but not necessarily, sanction the definitive break with the splinterists led by Renzi, the Democratic Party is looking for its government partner behind the m5s match that has recently opened. The majority of the party, the Nazareno insists, for now "does not intend to participate in an internal account"
Modification 2:	Dato il seguente titolo di un articolo di giornale scrivi lárticolo. ### Titolo: "Nel Pd niente conte interne". Ma su un annuncio già I nomi dei pro M5S ### Articolo: Alla vigilia al voto che potrebbe, ma non necessariamente, sancire la separazione con I dem guidati da Renzi, il pd cerca di chiudere la partita che tra poco si apre. La maggioranza del partito, insistono al Nazareno, è ferma sulla linea del "non cedere ad un conto interne"
Translation:	Given the following title of a newspaper article, write the article. ### Title:"No internal counts in the Democratic Party". But on an announcement already the names of the M5S pros ### Article: On the eve of the vote that could, but not necessarily, sanction the separation with the democrats, the Democratic Party tries to close the game that will soon open . The majority of the party, the Nazareno insists, is fixed on the idea of "not giving in to an internal account"

Table 12: Examples of Italian text modifications from IT5 models, used to compute the DetectGPT score. We show the translation and two modifications, in red the chunks that have been replaced. Depending on the length the modifications can alter the meaning of the original sentences, however they work to normalize a sentence likelihood and compute *DetectGPT*.

D.1 Modifications from IT5

The modification from IT5 to compute the DetectGPT score don't necessarily need to have a meaning as they are only used to normalize the log-likelihood values, however, it is interesting to note that often they don't disrupt the sentence meaning, and that they can create new sentences, although with similar meaning, that can themselves be useful as synthetic data. We release 600k of these sentences that are half modifications of the original news and half modifications of the synthetic texts generated by our fine-tuned models.



(a) GPT2 (b) Llama 13B

Figure 6: ROC curve for DetectGPT and log-likelihood. In (a) for the GPT2 model over 100 samples from xsum (coherent with (Mitchell et al., 2023)), in (b) for llama13b model over 100 samples from xsum.

	DetectGPT			*log-likelihood*
	t5-base	t5-3b	t5-11b	
Llama 13b	66%	71%	75%	78%
Llama 65b	-	62%	66%	70%

Table 13: AUROC achieved by *DetectGPT* (varying bootstrapping model) and *log-likelihood* on the xsum data-set for Llama 13B and Llama 65B.

Temperature	*DetectGPT*	*log-likelihood*
0.6	48%	86%
0.8	63%	73%
1.0	77%	52%

Table 14: AUROC achieved by *DetectGPT* and *log-likelihood* on Llama 13B varying the temperature used while generating the synthetic sentences.

generator	llama-2-7b_xsum		mistral_xsum	
	dGPT	llh	dGPT	llh
llama-2-7b	0.82	0.63	0.65	0.52
llama-2-7b_xsum	0.86	0.69	0.65	0.53
(995 samples) llama-2-7b_xsum	0.89	0.75	0.65	0.54
Mistral-7B-v0.1	0.59	0.41	0.77	0.59
mistral_xsum	0.57	0.44	0.85	0.75
(995 samples) mistral_xsum	0.53	0.43	0.95	0.92

Table 15: The AUROC achieved by all the models (rows) at different levels of fine-tuning, from pretrained only to fine-tuned on the full dataset. In **bold** the higher AUROC in each column.

E Generalizability to English

E.1 Synthetic Text Detection Based on Token Likelihoods

To confirm that our setup in subsection 5.1 is comparable to the previous efforts, we start by replicating a core result by Mitchell et al. (2023): on xsum data (Narayan et al., 2018), and using different versions of *t5* (Raffel et al., 2020) DetectGPT outperforms the *log-likelihood* in detecting GPT2 text (Radford et al., 2019) (see Figure 6a).

We apply the same methodology to sentences obtained using Llama 13B and Llama 65B. For Llama, we were unable to get *DetectGPT* to achieve a higher AUROC than the *log-likelihood*. We believe this to be due to the stronger performance of Llama compared to the t5 model used to generate new sentences (Figure 6b). This suggests that English text generated by Llama is harder to detect.

To test this hypothesis we measure the importance of the bootstrap model in this case. Table 13 shows the AUROC of *DetectGPT* depending on the bootstrap model, a larger t5 model leads to higher AUROC.

We repeat the experiment with both DetectGPT and log-likelihood at various temperature settings (0.6, 0.8, 1.0), and we find a strong sensitivity to this hyper-parameter, which merits further investigation.

To establish a fair comparison with DetectGPT while testing Llama, we perform an ablation study based on varying the temperature used in generation.

Table 14 shows different AUROC values for different temperatures. It appears that there is a strong sensitivity of the detection methodologies to this hyper-parameter, which merits further investigation in future work. The value 0.8 where *DetectGPT* and *log-likelihood* are more aligned, is also the value reported in the Llama repository.

E.2 Detecting CFMs with Proxy Models

We repeat the experiments done for the CHANGE-it dataset also for the XSUM dataset (Narayan et al., 2018).

That is we fine-tuned *llama-2-7b* and *mistral* on the full XSUM training set. We generate 2 datasets with 1k synthetic texts generated with each of the fine-tuned models and 1k samples from the xsum test set (less than for CHANGE-it to limit compute costs). Then we also fine-tune these two models on a small subset of the trainig set, 995 samples and finally we compute *DetectGPT* achieved by these 6 models, *llama-2-7b* and *mistral* pre-trained, *llama-2-7b_xsum_995* and *mistral_xsum_995* fine-tuned on 995 samples and *llama-2-7b_xsum* and

mistral_xsum fine-tuned on the xsum dataset.

Table 15 shows the results that closely match those for CHANGE-it shown in Table 2, namely that fine-tuning also on a small set of the same domain leads to high AUROC while if models come from different pre-trained ones the performance is low.

F Prompting details

Figure 7 shows the exact prompt that was used for generating the synthetic texts in section 6. In the prompt, we retain a few initial tokens of the original article, ensuring that the prompt never exceeded 30 words in total.

```
"""Given the following article title, generate the article.
### Title:
{title}
### Article:
{article}"""
```

Figure 7: Prompt used for generating news-like texts in section 6.

G Receiver Operating Characteristic (ROC) Curves

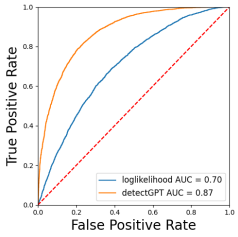
The receiver operating curve is a measure to understand how setting a threshold over a score would influence the true positive rate (TPR, the numer of instances marked as positive divided by the total number of positives) and the false positive rate (FPR, the number of false positive divided by the number of all negatives). That means how setting all instances with a score above a certain threshold as positive would influence the number of true positive and false positive.

Given that when we set as a threshold the maximum value of a score all instances are classified as negative TPR equals 0 and FPR equals 0, viceversa when it is set to the minimum TPR equals 1 and FPR equals 1. The ROC curve is a plot of the TPR against the FPR for different thresholds.

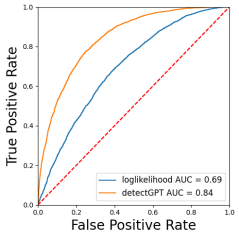
To show further details about the proxy models AUROC values shown in Table 2. Figure 8 shows the ROC curves for the case when the proxy models come from the same pre-trained one.

We can see tha computing the AUROC over 10k examples leads to a smooth curve and that the shape is very consistent across different amounts of fine-tuning futher validating the strngth of proxy models.

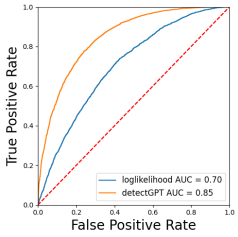
Llama-2-13b_it



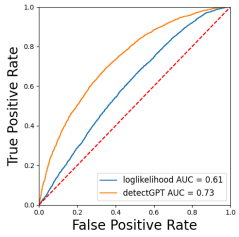
(a) *llama-2-13b_it*



(b) *llama-2-13b_it_3981*

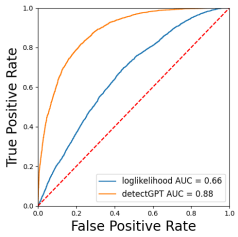


(c) *llama-2-13b_it_7862*

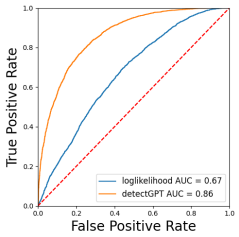


(d) *llama-2-13b*

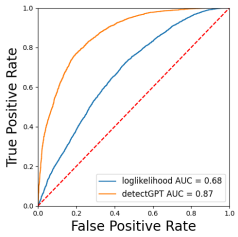
Llama-2-7b_it



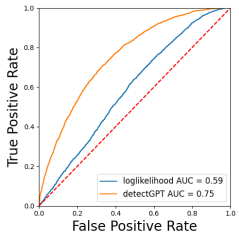
(e) *llama-2-7b_it*



(f) *llama-2-7b_it_3981*

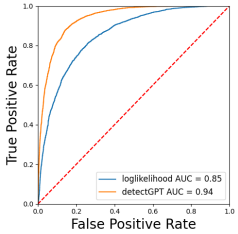


(g) *llama-2-7b_it_7862*

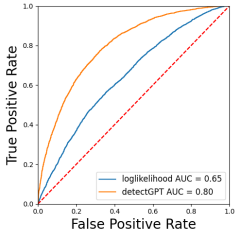


(h) *llama-2-7b*

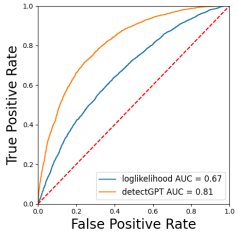
Mistral_it



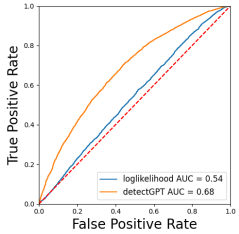
(i) *llama-2-13b_it*



(j) *llama-2-13b_it_3981*



(k) *llama-2-13b_it_7862*



(l) *mistral*

Figure 8: The ROC curves for the proxy models when evaluated on data generated by models fine-tuned starting from the same pre-trained model.

H Computational Costs

The fine-tuning run of our 'CFM' Llama (§3) lasted 5 days, as we wasted approximately 2 days due to exploding loss. Experiments were conducted using a private infrastructure, which has a carbon efficiency of 0.432 $kgCO_2eq/kWh$. A cumulative of 2,304 hours of computation was performed, and total emissions are estimated to be 298.6 $kgCO_2eq$.

The fine-tuning of *llama-2-7b_it*, *llama-2-13b_it* and *mistral_it* on CHANGE-it took 64 GPU hours each on A100 64Gb GPUs. With the costly synthetic data generation, all together resulted in approximately 2000 GPU hours. Experiments were conducted using the LEONARDO cluster, which has a carbon efficiency of 0.432 $kgCO_2eq/kWh$. A cumulative of 2000 hours of computation was performed on hardware similar to A100 PCIe 40/80GB (TDP of 250W). Total emissions are estimated to be 216 $kgCO_2eq$ (Luccioni et al., 2019).

Thus, we estimate that the total emissions for experiments in this study amount to approximately 515 $kgCO_2eq$.