PRISMAI: An Environment for AI-generated Text Recognition

Anonymous ACL submission

Abstract

We introduce PRISMAI, an environment for the automatic detection of AI-generated text. Our contributions are threefold: Firstly, we release the largest AI-detection dataset to date, comprising 537.588 human-written and AI-generated documents in both English and German across seven domains, including scientific writing, weblogs, parliamentary speeches, legal court cases, classic literature, news articles, and student essays, synthesized using state-of-the-art models. Secondly, we introduce LUMINAR, a CNN-based model for the automatic detection of AI-generated texts. Our experiments show that by leveraging the hidden states of an LLM to derive intermediate likelihoods, our model, despite having a small footprint, can outperform other likelihood-backed baselines significantly while demonstrating strong generalization capabilities in out-of-domain and out-of-language scenarios. Thirdly, we unify existing datasets into a common corpus called AIGT-WORLD and make it accessible through a publicly available web-based corpus explorer, which facilitates searching, reading, visualizing, and interacting with the underlying data. By doing so, we aim to elevate research in this area, expand the field to include non-English texts, propose new models, and unify existing efforts to build toward a common dataset and objective.

1 Introduction

With the advent of the Transformer architecture (Vaswani et al., 2017), Large Language Models (LLMs) have been widely adopted in various aspects of daily life (Veselovsky et al., 2023; Murakami et al., 2023; Jiang et al., 2024), owing to their ability to generate high-quality text (OpenAI et al., 2024; Team et al., 2024a), in some cases even surpassing human writing (Gómez-Rodríguez and Williams, 2023). This has led to a massive influx of AI-generated text on the internet, with recent studies indicating a 57.3% increase on mainstream websites and an 474% increase on misinformation sites (Hanley and Durumeric, 2023), alongside a rise in plagiarism cases (Bisi et al., 2023; Elali and Rachid, 2023; Pudasaini et al., 2024). This issue was further exacerbated by Lu et al. (2024), who introduced the first fully automated *AI Scientist* — a system composed of AI agents capable of generating entire scientific papers, from idea generation to experimentation and final manuscript writing.

Consequently, concerns have emerged regarding the long-term implications of AI-generated text, not only for information reliability and content originality but also for the broader digital ecosystem (Shen and Zhang, 2024; Wang and Lu, 2025). For instance, Shumailov et al. (2024) have shown that recursively training language models on AI-generated text (AIGT) leads to irreversible degradation in model quality, a phenomenon known as *model collapse*, where later-generation models experience catastrophic forgetting and a decline in linguistic coherence.

To address these issues, research on the automatic detection of AI-generated text has emerged as a crucial field of study. Herein, the collection of datasets (Yu et al., 2023; Su et al., 2023b; Li et al., 2024) and the development of AI-detection models (Wang et al., 2023; Verma et al., 2024) have become the two primary objectives. With PRISMAI, we contribute to this research by, first, releasing the largest AI-detection dataset to date (see Table 1), and second, introducing LUMINAR a novel AI-written text detection model that uses Convolutional Neural Networks on a likelihood-based feature space to detect AI-generated snippets within documents. Thirdly, as part of the PRISMAI framework, we release everything as open-source[1], unify existing datasets into a single corpus called

[1] https://anonymous.4open.science/r/PrismAI-ACL-824C/

AIGT-WORLD, and provide an interactive corpus explorer via a publicly accessible web portal[2] that enables users to explore both human-written and AI-generated content through semantic searches and intuitive visualizations (screenshots of the portal are provided by Figure 5 in the appendix).

2 Related Work

We consider two parts of related work. First, we discuss models that have been developed for the automatic detection of various forms of AI-generated text. Second, we examine the most recent datasets collected in connection with these efforts.

2.1 Models

Several approaches focus on the automatic detection of machine-written texts. One of them is based on linguistic analysis, including n-gram frequencies (Badaskar et al., 2008), entropy (Gehrmann et al., 2019), and log-likelihood-based methods, where the latter has led to some of the most prominent models to date: *DetectGPT* (Mitchell et al., 2023), and its successors *DetectLLM* (Su et al., 2023a) and *Fast-DetectGPT* (Bao et al., 2024). *DetectGPT* relies on the *perturbation* of input sequences and its effect on the log-likelihood of predictions, while *Fast-DetectGPT* relies on the sampling of a large number of alternative tokens from the conditional probability distribution produced for each token in an input sequence by an LLM. Therefore, their main limitation is their white-box nature, requiring access to model logits and prediction distributions, which is often unavailable, especially when using APIs or services such as ChatGPT.

A second approach focuses on *watermarking* LLM-generated text, as done by Kirchenbauer et al. (2024) and Lee et al. (2024), the latter primarily addressing the detection of machine-generated code. In this approach, watermarked text can be generated using standard language models without requiring retraining, while the misclassification of human-generated text as machine-generated remains statistically unlikely (Kirchenbauer et al., 2024). Although this research direction is actively explored (Liu et al., 2024; Zhao et al., 2024; Dathathri et al., 2024) and yields promising results, it is not relevant to our scenario, as we focus on detecting AI-generated text in a fully black-box scenario, where AI-generated content may be deliberately hidden, and only the text is provided.

Since we focus on black-box scenarios where access to the generating LLM or potential watermarks is unavailable, leveraging open-source LLMs for continuation (Yang et al., 2023), rewriting (Mao et al., 2024), or paraphrasing (Quidwai et al., 2023) of texts as part of feature engineering or similarity calculations is a widely adopted approach. Furthermore, labeled as state-of-the-art, *Ghostbuster* (Verma et al., 2024) utilizes a series of weaker language models to obtain token probabilities, which are then used to perform a structured search across model combinations. A linear classifier is subsequently trained to distinguish between AI- and human-written texts. Finally, given the abundance of document-level classification models, Wang et al. (2023) introduced *SeqXGPT* for sentence-level classification. This approach leverages log probabilities from white-box LLMs, which are processed through convolutional and self-attention networks.

2.2 Datasets

Several datasets have been compiled for the detection of AI-generated text (Wu et al., 2025). *HC3* (Guo et al., 2023) and *HC3 Plus* (Su et al., 2023b) are among the first datasets of this kind. They consist of both human-written and AI-generated texts in English and Chinese, making them unique due to the rarity of bilingual data, with a focus on news, translations and Q&A texts. The *CHEAT* dataset (Yu et al., 2023) was created to detect plagiarism in the scientific field. It includes human-written abstracts alongside their ChatGPT-generated summaries, and so-called *fusion* texts, as a mix of human and AI-generated content. One of the largest datasets, *OpenLLMText* (Su et al., 2023a), comprises over 340 000 texts derived from various AI models, with a primary focus on web texts. To detect AI-generated content in academic contexts, such as classroom exercises, Liu et al. (2023) propose *ArguGPT*, a collection of 4 000 argumentative essays generated by seven different GPT models, but with no human-written counterparts. A broader dataset covering a wider range of domains has been introduced by Pu et al. (2022) in the form of *DeepfakeTextDetect*; it includes news articles, stories, scientific texts and more, generated using a variety of models. This dataset is not open-source, as access requires a Google Form submission, which grants the authors the right to withdraw their data at any time and prohibits further distribution.

[2] *anonymized* - link won't be available during review

Table 1: The datasets unified within AIGT-WORLD and their statistics, including our PRISMAI-dataset. The notation ⟨*models*⟩ $\Longrightarrow$ ⟨*domains*⟩ indicates that the set of *models* was applied to the given set of *domains* to generate AI counterparts. The *MAGE* dataset uses 27 LLMs, which are variations of the listed base models. The word and sentence counts for the Chinese part of the *HC3-Plus* dataset may not be representative.

AIGT-WORLD						
Dataset	Languages	Label	Documents	Words	Sentences	Size (MB)
SeqXGPT-Bench (Wang et al., 2023)	English	AI	29 904	6 267 328	358 973	36.07
		Human	6 000	1 424 983	81 801	8.04
⟨ GPT-2, GPT-J, GPT-NEO, LLaMA, GPT-3.5-turbo ⟩ $\Longrightarrow$ ⟨ news, social media posts, web texts, scientific articles, technical documentation ⟩						
CHEAT (Yu et al., 2023)	English	AI	30 790	4 490 024	246 586	29.97
		Human	15 395	2 458 466	133 384	15.92
		HU × AI	4 514	649 925	35 480	4.28
⟨ GPT-3.5-turbo ⟩ $\Longrightarrow$ ⟨ scholarly literature ⟩						
Ghostbuster (Verma et al., 2024)	English	AI	12 000	6 010 131	385 358	38.42
		Human	2 000	1 192 091	86 977	6.76
⟨ GPT-3.5-turbo, Claude ⟩ $\Longrightarrow$ ⟨ creative writing, news, student essays ⟩						
HC3-Plus (Su et al., 2023b)	English Chinese	AI	144 582	8 026 922	540 497	54.31
		Human	178 939	9 643 121	722 092	56.77
⟨ GPT-3.5-turbo ⟩ $\Longrightarrow$ ⟨ news, translations, question answering ⟩						
OpenLLMText (Su et al., 2023a)	English	AI	275 546	92 334 033	5 277 345	546.19
		Human	68 984	38 603 989	2 264 905	229.03
⟨ GPT-3.5, PaLM, LLaMA-7B, GPT2-1B ⟩ $\Longrightarrow$ ⟨ web texts ⟩						
MAGE (Li et al., 2024)	English	AI	281 824	61 590 290	4 532 144	343.97
		Human	150 858	30 093 650	2 325 683	163.52
⟨ OpenAI GPT, LLaMA, GLM-130B, FLAN-T5, OPT, BigScience, EleutherAI ⟩ $\Longrightarrow$ ⟨ opinion statements, reviews, news, question answering, story generation, reasoning, wikipedia, scientific writing ⟩						
PrismAI	English German	AI	404 066	132 653 218	8 016 429	834.57
		Human	154 978	626 961 931	40 229 029	3 408.67
⟨ GPT-4-turbo, GPT-4o-mini, o3-mini, Nemotron, Gemma2-9b, DeepSeek-r1:1.5b & 32b, phi3-3.8b ⟩ $\Longrightarrow$ ⟨ scientific writing, weblogs, parliamentary speeches (English & German), news (English & German), legal court cases, classic literature (English & German), student essays ⟩						

Recently, Li et al. (2024) proposed *MAGE*, a testbed with the widest variety of domains to date, including opinion statements, story generation, scientific writing and more. The dataset, which consists entirely of English text, uses 27 different LLMs and contains over 400 000 samples, the majority of which are AI-generated. In order to train their model, *Ghostbuster*, Verma et al. (2024) collected three new datasets covering the domains of creative writing, news, and student essays. They used *gpt-3.5-turbo* (Brown et al., 2020) to generate AI counterparts, creating 14 000 texts. While current AI-generated text detection focuses mainly on document-level classification, Wang et al. (2023) synthesized a sentence-level detection dataset. It includes sources such as news articles, social media posts, and technical documentation, and is based on the *SnifferBench* (Li et al., 2023) dataset.

3 Data

While the datasets discussed in Section 2.2 provide a solid foundation, there are still three key gaps. First, the dominance of English texts fails to reflect linguistic diversity. Second, the vast majority of AI-generated text is produced by models that are outdated by today's standards, including GPT-2, GPT-3.5, and LLaMA (Touvron et al., 2023), raising concerns about their relevance in terms of both performance and text generation. Third, because the focus is primarily on document-level detection, they lack fusion texts, specifically chunk-based AI snippets embedded in human-written content.

To fill these gaps, we introduce the PRISMAI-dataset. It consists of (A): English news articles collected from **CNN-DailyMail** (See et al., 2017) and German news articles from **Spiegel Online**,

(B) English scientific articles harvested by scraping **arXiv** papers and extracting their full PDF content via `PyMuPDF`, (C) English web-blogs from **blogger.com** (Schler et al., 2006), (D) German parliamentary speeches provided by **Anonymous**[3] and English speeches from the **House of Commons** archive (Blumenau, 2021), (E) English legal cases from the **European Court of Human Rights** (Chalkidis et al., 2021), (F) English student essays written by 6th–12th grade students, collected within a kaggle competition (King et al., 2023) and built upon Crossley et al. (2024), and (G) English and German classic literature texts scraped from Project Gutenberg (Gutenberg, n.d.). For generating the AI texts, we employ state-of-the-art models with a focus on publicly accessible and popular LLMs, including **Gemma 2** (9B) (Team et al., 2024b), **Phi-3** (3.8B) (Abdin et al., 2024), **DeepSeek-R1** (1.5B | 32B) (DeepSeek-AI et al., 2025), **Nemotron** (70B) (Nvidia et al., 2024) and OpenAIs **GPT-4-Turbo**, **GPT-4o-Mini** (OpenAI, 2024) and **o3-Mini** (OpenAI, 2025).

We also address the lack of chunk-based AI snippets that resemble *fusion* texts (Yu et al., 2023). Since AIGT is often embedded in human-written texts, we categorize our dataset into *chunked* and *fulltext*. To generate *chunked* AI texts, we divide human-written texts into segments and replace up to 50% with a contiguous selection of masked chunks. An LLM then reconstructs the missing content based on the surrounding human-authored text. For *fulltext*, we prompt the LLM to extract key information from a human-written text in a structured format. This includes identifying the language, reconstructing contextualized scenarios, determining the topic, and analyzing linguistic style. Based on *few-shot prompting*, we provide an example alongside the desired output. Using the extracted information, we then prompt the LLM in a ghost-writing scenario to generate a new text based on the details provided, including a specified word count (for the details see Appendix **??**). Finally, we add the *fulltext* and *chunked* AI-generated variants of all human-written texts to PRISMAI.

4 LUMINAR

We now introduce our model for AI-generated text detection, LUMINAR, a CNN-based text classification model (LeCun et al., 1989) that builds upon the approaches outlined in Section 2. We discuss the features used to power LUMINAR and its architecture, following by an evaluation of its performance on a subset of PRISMAI.

4.1 Features

Inspired by *DetectLLM* (Su et al., 2023a) and *Fast-DetectGPT* (Bao et al., 2024), we investigate the likelihoods of common LLMs to derive features for LUMINAR. Let $p_\theta(x)$ be the probability of any given token x following a sequence of preceding tokens parametrized by the models' weights θ (Radford et al., 2019):

$$p_\theta(x) = \prod_{i=1}^{n} p(x_i | x_1, \dots, x_{i-1})$$

Contemporary LLMs approximate these probabilities by passing encoded inputs, usually consisting of token and position embeddings, through layers of identical *Transformer* blocks TF_θ (Vaswani et al., 2017; cf. Radford et al., 2019; Dubey et al., 2024; Team et al., 2024b) forming hidden states HS between each layer. The output of the last *Transformer* layer is then passed through a *Language Modeling Head* (LM head, LMH), which acts as a projection layer from hidden states to vocabulary space and produces the logits. Given an input encoding method Enc_θ and a tokenized input sequence $\mathbf{x}$, we can formalize these operations (slightly simplified; omitting model specific operations, activation functions, etc.) as:

$$\begin{aligned} HS_0 &= Enc_\theta(\mathbf{x}) \\ HS_i &= TF_{\theta_i}(HS_{i-1}) \\ logits &= \mathrm{LMH}_\theta(HS_H) \end{aligned}$$

where there exists a TF_{θ_i} and a HS_i for each $i \in \{1, \dots, H\}$ for an H layer model. Applying $SoftMax(logits)$ yields the likelihood vector L containing value for each token conditioned on its preceding tokens.

Previous work uses pre-trained white-box LLMs to determine the likelihood of a given, potentially permuted token for a given text, and to infer judgments about human authorship. We use the same method but also include **Intermediate Likelihoods** (**IL**), i.e., a set of likelihoods obtained by the intermediate hidden states of an LLM. We discuss IL and a number of features we considered, including the regular *Likelihood* (**L**), the top-k token's likelihoods in terms of the *top-k Likelihood Likelihood*

[3] Published at the Jurix 2023 conference

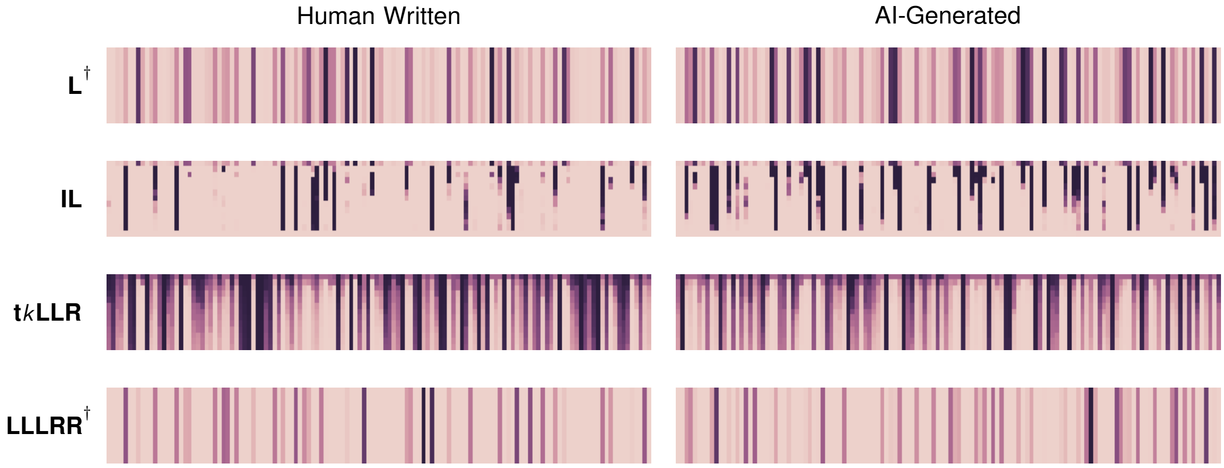


Figure 1: Visualization of four feature types for the first 128 tokens of a *CNN-DailyNews* article and its AI-generated counterpart (created using `gpt-4o-mini`), utilizing `gpt2`'s hidden states. Darker shades code higher values. Features from top to bottom: **L**ikelihoods; **I**ntermediate **L**ikelihoods; **t**op-**k** **L**ikelihood **L**ikelihood **R**atio. **L**og-**L**ikelihood **L**og-**R**ank **R**atio; †The features are 1D but are shown as 2D for clarity; all other features have 13 dimensions each.

Ratio (**t*k*LLR** and its inverse **Lt*k*LR**), and the *Log-Likelihood Log-Rank Ratio* (**LLLRR**, derived from LLR, cf. Su et al., 2023a).

Intermediate Likelihoods As each *Transformer* layer has the same input/output characteristics we can simply pass each hidden state through the LM head to calculate the intermediate likelihoods L_i:

$$logits_i = \mathsf{LMH}_\theta(HS_i)$$
$$L_i = \mathsf{SoftMax}(logits_i)$$

Let now $p_i(x_j) = L_{ij}$ be the intermediate likelihood of token x_j for layer i and $p(x_j) = p_H(x_j)$ the "regular" likelihood. Thus our first feature, the *Likelihoods*, are given as vector $\mathbf{L} = \begin{pmatrix} p(x_1) & \dots & p(x_n) \end{pmatrix}$. Using the likelihoods of the intermediate hidden states, we define our second feature, the *Intermediate Likelihoods*, as a matrix $\mathbf{IL}$ where the rows represent the intermediate likelihood from one layer and the columns represent each token as such that:

$$\mathbf{IL}_{ij} = p_i(x_j)$$

t*k*LLR Let $\mathsf{top_k}(j)$ be the k-highest likelihood at position j. Thus the *top-k Likelihood Likelihood Ratio* is given as matrix a with values

$$\mathbf{t}k\mathbf{LLR}_{kj} = \frac{\mathsf{top_k}(j)}{\mathsf{top_k}(j) + p(x_j) + \epsilon}$$

where $k \in \{1, \dots K\}$ is a variable sized hyperparameter. We may also consider the inverse ratio, **Lt*k*LR**, where $p(x_j)$ is placed in the numerator.

LLLRR By calculating the LLR for each token x_j with likelihood $p(x_j)$ and the corresponding 1-indexed rank $r(x_j)$ separately, we define our *Log-Likelihood Log-Rank Ratio* feature as vector

$$\mathbf{LLLRR}_j = -\frac{\log\left(p(x_j)\right) + \epsilon}{\log\left(r(x_j)\right) + \epsilon}$$

adding a small constant ϵ to avoid $\log(0)$ and $\frac{\cdot}{0}$, respectively. [4]

Example Figure 1 shows a set of features for two texts from the *CNN-DailyMail* dataset, with the human-written text features' on the left and the AI-generated text features' on the right, for the first 128 tokens of each document generated with `gpt2`, where darker colors code higher values. The first row shows the *Likelihood* features, where we can see a small set of characteristic high-likelihood areas in the first fifth of the features corresponding to the tokenized text

```
at the christ ␣ening of Prince George
```

where the second token '`the`' has a likelihood of 0.3 while the fourth token '`␣ening`' has a likelihood of 0.7 (␣ indicating a continuation token). Below that, we see the *Intermediate Likelihoods*, where the upper row pertains to the likelihood in the last layer and thus matches the *Likelihood* in the first row. We can clearly see the 1.0 likelihood in the lower layers (first vertical dark line) for

[4] While SoftMax likelihoods *should* never be zero, we encountered several cases where numerical instability or loss of precision during data format transitions led to zero likelihoods.

'the', but a low likelihood 0.0 for '␣ening'.[5] Note that there are only few such patterns in the human-written sample, while the AI-generated sample shows a large number of high-likelihood columns for the lower layers of the model.

The third row shows the *top-k Likelihood Likelihood Ratio*. The likelihood acts as a scaling factor against the top-k tokens, highlighting areas of ambiguity by significantly increasing the top-k likelihood values when the likelihood is small or equal to them.

The fourth row shows the *Log-Likelihood Log-Rank Ratio*, the metric used in perturbation-free *DetectLLM-LLR* approach of Su et al. (2023a). It is the only feature that is not restricted to the unit range $[0, 1]$.

4.2 Document-Level Features

In the general case, we treat AI text detection as a *text classification task*. To this end, we generate text-level features from an input text by first computing the hidden states with a simple forward pass of the tokenized text through an LLM. Then, we compute all *Intermediate Likelihoods* by passing each hidden state through the LM head. This results in a two-dimensional vector $\vec{feat}_D \in \mathbb{R}^{L \times H}$, where L is the length of the tokenized document and H is the number of hidden states (e.g. 13 for `gpt2` or 17 for `Llama3.2-1B`). We experimented with different sampling methods and found that taking the features or randomly selecting strided slices performed the best. To obtain a vector of fixed length from texts of variable size L, we sample k slices of size s from $\vec{feat}_D$ and concatenate them into a single feature vector $\vec{feat}$. If a text is too short to sample sufficient slices, we fall back on the first $k \cdot s$ features and apply zero padding to the right if necessary.

4.3 Model

The LUMINAR text classification model is backed by a CNN with a linear classifier (see Figure 2). Depending on the operative feature selection method of Section 4.1, we first concatenate the selected feature slices along the sequence length dimension. We experimented with different architectures and found that multiple 1D convolution layers performed best across feature variants. Thus, we pass our two-dimensional features as single vectors with multiple channels into the CNN. The CNN consists of multiple layers with `LeakyReLU` activation functions followed by a linear binary classifier with an optional projection layer.

[5] See Section A.2 for the full text of this example.

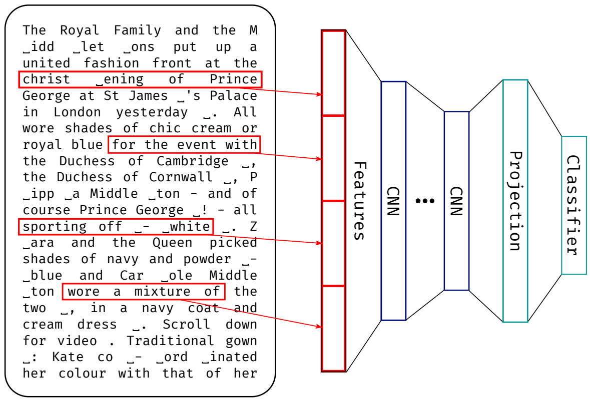


Figure 2: LUMINAR Model Overview.

5 Experiments

During training, we always group human-written documents with their AI-written counterparts before splitting the dataset into training, validation and test sets, so that these paired texts do not appear in different splits. We apply five-fold cross validation with 70 % of the document groups in training, 10 % in validation, and 20 % test splits. We test the performance of our models both *in-domain* and *out-of-domain* (OOD) by training models on one domain or a combination of all domains and testing performance on all domains, and by training a model on all but one domain and testing performance on the held-out domain. We include two German corpora (news and parliamentary speeches, cf. Table 1) in our data to evaluate the cross-lingual performance of the model.

To train the models, we create sized datasets of human texts and individual AI-generated counterparts as well as the combination of multiple agents' texts. To this end, we consider a subset of 1500 documents for each domain with synthetic texts generated by `GPT-4o-Mini`. To generate our features, we use `GPT-2` and truncate documents that are longer than 1024 tokens to the models maximum context size. We considered using larger LLMs but found the performance of our models with features from `GPT-2` (with 124 million parameters) to be sufficient already. Our experiments were conducted using `PyTorch` (Paszke et al., 2019), `Lightning` (Falcon and team) and `transformers` (Wolf et al., 2020). To ensure reproducibility, we run our training in deterministic mode and record all used hyperparameters with the results.

Table 2: Results for LUMINAR and likelihood-based baselines in the in-domain setting.

Domain	First		Random		LLR		Fast-DetectGPT	
	AUROC	F_1	AUROC	F_1	AUROC	F_1	AUROC	F_1
Web Blogs	**1.000**	**1.000**	0.996	0.968	0.470	0.493	0.369	0.587
Essays	**0.971**	**0.895**	0.979	0.895	0.846	0.779	0.925	0.833
CNN	**0.995**	**0.947**	0.981	0.912	0.942	0.876	0.972	0.915
ECHR	**0.996**	**0.953**	0.991	0.934	0.916	0.851	0.820	0.744
HoC	**0.990**	**0.938**	0.973	0.881	0.856	0.829	0.831	0.809
arXiv	**0.995**	**0.977**	0.984	0.933	0.965	0.937	0.866	0.847
Gutenberg	**0.978**	**0.910**	0.971	0.912	0.859	0.767	0.907	0.874
Bundestag_{de}	**0.991**	**0.963**	0.963	0.892	0.846	0.787	0.799	0.748
Spiegel_{de}	**0.968**	**0.936**	0.930	0.845	0.868	0.804	0.782	0.701

5.1 Training

The LUMINAR models discussed below were all trained as follows, unless otherwise noted. We train the models for up to 25 epochs using `AdamW` (Loshchilov and Hutter, 2019) with `lr = 0.0001` and a linear learning rate scheduler with a warmup period of one epoch, employing early-stopping conditioned on the validation loss after three consecutive epochs without improvement. In case of **Random** features, we concatenate likelihoods by from four randomly selected slices of size 64 with stride 16, treating the second dimension as input channels. The CNN features five `Conv1D` layers with `stride=1` and `(channels, kernel size)` of `[(64,5),(128,3),(128,3),(128,3),(64,3)]`.

5.2 Results

Table 2 shows the results of our models and selected likelihood-based baselines for each domain individually. F_1-Scores for *DetectLLM*'s LLR and *Fast-DetectGPT* were obtained by calculating metrics for the whole domain dataset and choosing a reasonable threshold by finding the middle point between the means of the distributions for each text class. The scores for our models are the averages of five-fold cross validation using both the *First* and *Random* feature slicing methods as outlined in Section 4.2. We use 0.5 as a fixed threshold to calculate the F_1-Score for our classifiers.

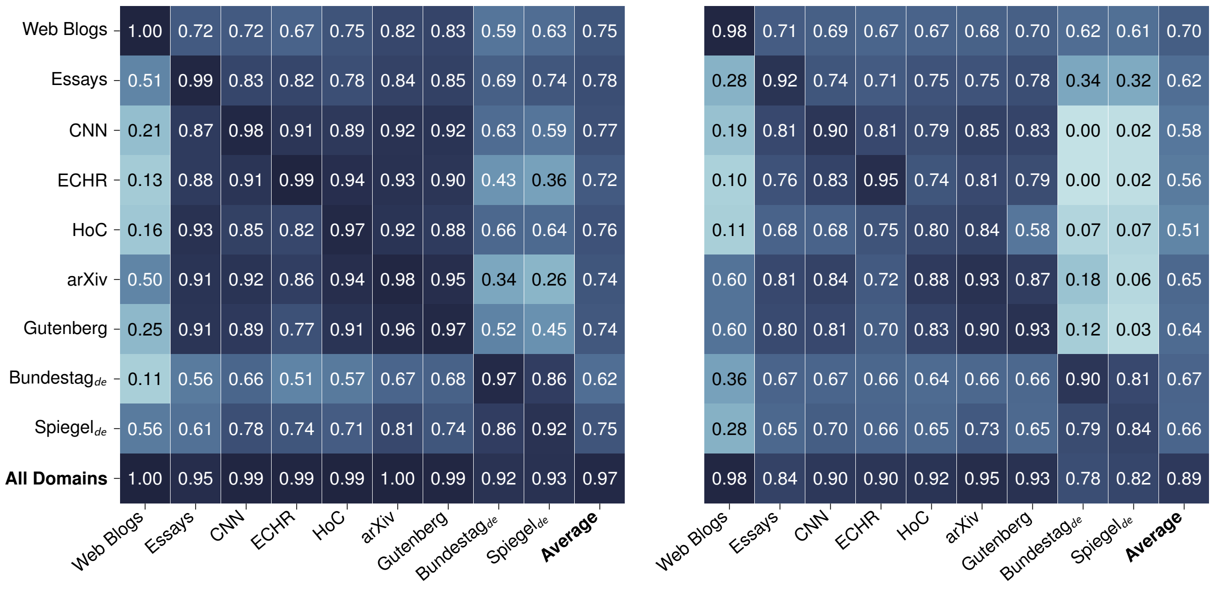
Our classifier outperforms the baselines on all domains, in part with a significant margin. This is especially apparent for the *Blog Authorship* domain, which contains many documents with highly irregular language (see Appendix A.2 for an example). The statistics-based baselines struggle with these irregularities, while our classifier trained on the first 256 tokens' intermediate likelihoods returns a perfect score. Our model also works very well in a cross-lingual setting, with F_1-scores of 0.963 and 0.936 for the two German subsets, *Bundestag* and *Spiegel Online*, respectively.

Table 3: LUMINAR out-of-domain results.

Domain	First		Random	
	AUROC	F_1	AUROC	F_1
Web Blogs	0.296	0.411	**0.446**	**0.474**
Essays	0.624	0.699	**0.763**	**0.732**
CNN	**0.973**	**0.908**	0.973	0.883
ECHR	**0.945**	**0.884**	0.940	0.853
HoC	**0.967**	**0.887**	0.959	0.853
arXiv	**0.993**	**0.912**	0.971	0.857
Gutenberg	0.917	0.846	**0.971**	**0.897**
Bundestag_{de}	**0.961**	**0.888**	0.795	0.723
Spiegel_{de}	0.506	0.175	**0.781**	**0.663**

Training on randomly sampled feature slices results in consistently worse performance in the in-domain setting. In the out-of-domain setting, that is training on all *but* the test domain, the results are more nuanced: the randomly sampled features yield stronger results for four out of nine domains. Most notably, in the $\textit{Spiegel}_{de}$ domain, the classifier trained on randomly sampled feature slices performs 0.488 points in F_1-score better than the one trained on the first set of features. In addition, the *Web Blogs* and *Essays* domains both fall well below 0.8 F_1-score, too, while the remaining domains retain their performance within ≤ 0.1 points.

On one hand, this highlights the need for diverse training data when training LUMINAR, but it also shows that the model has strong generalization capabilities. To further explore these, we conducted



(a) AUROC (b) F_1-Score

Figure 3: Heatmap of cross-domain evaluation scores of LUMINAR trained on randomly sampled `GPT-2` *Intermediate Likelihood* feature slices on human-written texts and texts generated using `GPT-4o-Mini`. The values shown are for in- and cross-domain settings for each training-test-domain combination, as well as a classifier trained on all domains (last row) and their respective averages (last column), where each row shows the domain which the model was trained on, while the columns pertain to test domain.

cross-domain experiments, training a model on a single as well as all domains and testing it against all domains individually. The heatmap in Figure 3 shows the results of the cross-domain evaluation.

The jointly trained model retains very good AUROC values across all domains while the F_1-score drops significantly, especially for the German domains. Despite being a purely English pre-trained LLM, features from `GPT-2` perform very well in the in-domain setting when trained on German texts. We hypothesize that using an LLM pre-trained on multilingual data would help bridge the gap in a multilingual application scenario for LUMINAR.

6 Discussion

Despite being a supervised method, our model with its default configuration is tiny, having only 676k parameters equating to a memory footprint of less than 3MB. Neither training nor inference of the model is computationally intensive and can be done on an average workstation with a consumer-grade GPU. However, for the baseline methods to achieve similar levels of performance, one would have to use much larger LLMs which outweighs the added computational load of training a CNN classifier by multiple levels of magnitude (cf. Table A.5).

When we were investigating related work to draw for a comparison, we noted that multiple publications only report the AUROC values for their models many times achieving values ≥ 0.99, while referring to it as a "commonly used metric" for AI-generated text detection research, e.g. Su et al. (2023a) referring to AUROC as a "commonly used to measure zero-shot detector performance, which considers the range of all possible thresholds". However, for any real-world use case, the AUROC has little value because a high AUROC does not imply high classification performance. As such, we follow Verma et al. (2024) by also calculating the F_1-score in all our experiments.

7 Conclusion

We introduced PRISMAI, an environment for AIGT detection featuring the largest multilingual corpus to date, along with LUMINAR, a CNN-based model for detecting AI-generated text. To advance and expand existing research, we have integrated our new PRISMAI dataset with prior efforts into a unified corpus, AIGT-WORLD, aiming to continuously enhance diversity across domains, models, and languages through a common data pool. Additionally, we introduce a first-of-its-kind corpus explorer for AIGT datasets, allowing users to visually search and interact with data rather than relying solely on programmatic access. Finally, we release everything as open source.

Limitations

Sequence Length

While CNNs do not depend on any particular input size, the *classifier* does. This implies limitations to the input sequence length, as some texts might be much shorter than the average training sample. While our experiments show good generalization for shorter sequences (cf. Table 6 for statistics on the sequence length of our dataset), we can leverage the translation-invariance of the convolution operation to create multiple classifiers backed by the same CNN, trained jointly on different feature sequence lengths, which we will explore in future work.

Calculating Intermediate Likelihoods

Calculating the intermediate likelihoods is no more or less computationally complex than calculating the regular likelihood. However, depending on the size of the model and the number of layers, this can lead to a significant memory overhead (i.e. `GPT-2`'s LM head makes up 31 % of the models parameters). We address this issue by transferring the hidden states to the CPU before calculating the intermediate likelihoods iteratively on the GPU.

Ethical Considerations

We trained LUMINAR on a diverse range of domains, including two languages, yet this does not encompass all writing styles and topics. As a result, texts resembling those seen during training tend to yield more confident predictions, whereas those from other domains exhibit greater variability. This includes languages as well. We have also seen differences in LUMINAR's behaviour depending on the length of the texts. In general, the predictions of LUMINAR are not legally binding and should not be considered as definitive or reliable in this context. Particularly in the context of automated plagiarism detection in school or university settings, LUMINAR's predictions should never be accepted without manual verification. We strongly oppose any unfiltered or fully automated integration of our model into such systems. While we cannot prevent misuse of our model, we still release it as part of the open-source PRISMAI environment for the benefit of research and broader accessibility.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Weizhu Chen, Yen-Chun Chen, Yi-Ling Chen, Hao Cheng, Parul Chopra, Xiyang Dai, Matthew Dixon, Ronen Eldan, Victor Fragoso, Jianfeng Gao, Mei Gao, Min Gao, Amit Garg, Allie Del Giorno, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Wenxiang Hu, Jamie Huynh, Dan Iter, Sam Ade Jacobs, Mojan Javaheripi, Xin Jin, Nikos Karampatziakis, Piero Kauffmann, Mahoud Khademi, Dongwoo Kim, Young Jin Kim, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Xihui Lin, Zeqi Lin, Ce Liu, Liyuan Liu, Mengchen Liu, Weishung Liu, Xiaodong Liu, Chong Luo, Piyush Madan, Ali Mahmoudzadeh, David Majercak, Matt Mazzola, Caio César Teodoro Mendes, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Liliang Ren, Gustavo de Rosa, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Yelong Shen, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Praneetha Vaddamanu, Chunyu Wang, Guanhua Wang, Lijuan Wang, Shuohang Wang, Xin Wang, Yu Wang, Rachel Ward, Wen Wen, Philipp Witte, Haiping Wu, Xiaoxia Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Jilong Xue, Sonali Yadav, Fan Yang, Jianwei Yang, Yifan Yang, Ziyi Yang, Donghan Yu, Lu Yuan, Chenruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *Preprint*, arXiv:2404.14219.

Sameer Badaskar, Sachin Agarwal, and Shilpa Arora. 2008. Identifying real or fake articles: Towards better language modeling. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-II*.

Guangsheng Bao, Yanbin Zhao, Zhiyang Teng, Linyi Yang, and Yue Zhang. 2024. Fast-detectgpt: Efficient zero-shot detection of machine-generated text via conditional probability curvature. *Preprint*, arXiv:2310.05130.

Théophile Bisi, Anthony Risser, Philippe Clavert, Henri Migaud, and Julien Dartus. 2023. What is the rate of text generated by artificial intelligence over a year of publication in orthopedics & traumatology: Surgery & research? analysis of 425 articles before versus after the launch of chatgpt in november 2022. *Orthopaedics & Traumatology: Surgery & Research*, 109(8):103694.

Jack Blumenau. 2021. Measuring political debate:

Responsiveness, influence, and rhetoric in parliamentary texts. Dataset. Sponsored by the Economic and Social Research Council, Grant reference: ES/N016297/1.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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 M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher 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. *Preprint*, arXiv:2005.14165.

Ilias Chalkidis, Manos Fergadiotis, Dimitrios Tsarapatsanis, Nikolaos Aletras, Ion Androutsopoulos, and Prodromos Malakasiotis. 2021. Paragraph-level rationale extraction through regularization: A case study on European court of human rights cases. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 226–241, Online. Association for Computational Linguistics.

S.A. Crossley, Y. Tian, P. Baffour, A. Franklin, M. Benner, and U. Boser. 2024. A large-scale corpus for assessing written argumentation: Persuade 2.0. *Assessing Writing*, 61:100865.

Sumanth Dathathri, Abigail See, Sumedh Ghaisas, Po-Sen Huang, Rob McAdam, Johannes Welbl, Vandana Bachani, Alex Kaskasoli, Robert Stanforth, Tatiana Matejovicova, Jamie Hayes, Nidhi Vyas, Majd Al Merey, Jonah Brown-Cohen, Rudy Bunel, Borja Balle, Taylan Cemgil, Zahra Ahmed, Kitty Stacpoole, Ilia Shumailov, Ciprian Baetu, Sven Gowal, Demis Hassabis, and Pushmeet Kohli. 2024. Scalable watermarking for identifying large language model outputs. *Nature*, 634(8035):818–823.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate

Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and et al. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Faisal R. Elali and Leena N. Rachid. 2023. Ai-generated research paper fabrication and plagiarism in the scientific community. *Patterns*, 4(3):100706.

William Falcon and The PyTorch Lightning team. Pytorch lightning.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander Rush. 2019. GLTR: Statistical detection and visualization of generated text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 111–116, Florence, Italy. Association for Computational Linguistics.

Carlos Gómez-Rodríguez and Paul Williams. 2023. A confederacy of models: a comprehensive evaluation of LLMs on creative writing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14504–14528, Singapore. Association for Computational Linguistics.

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. *Preprint*, arXiv:2301.07597.

Project Gutenberg. n.d. Project gutenberg. Retrieved February 21, 2016, from `https://www.gutenberg.org`.

Hans W. A. Hanley and Zakir Durumeric. 2023. Machine-made media: Monitoring the mobilization of machine-generated articles on misinformation and mainstream news websites. In *International Conference on Web and Social Media*.

Juyong Jiang, Fan Wang, Jiasi Shen, Sungju Kim, and Sunghun Kim. 2024. A survey on large language models for code generation. *Preprint*, arXiv:2406.00515.

Jules King, Perpetual Baffour, Scott Crossley, Ryan Holbrook, and Maggie Demkin. 2023. Llm - detect ai generated text. `https://kaggle.com/competitions/llm-detect-ai-generated-text`. Kaggle.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. 2024. A watermark for large language models. *Preprint*, arXiv:2301.10226.

Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. 1989. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551.

Taehyun Lee, Seokhee Hong, Jaewoo Ahn, Ilgee Hong, Hwaran Lee, Sangdoo Yun, Jamin Shin, and Gunhee Kim. 2024. Who wrote this code? watermarking for code generation. *Preprint*, arXiv:2305.15060.

Linyang Li, Pengyu Wang, Ke Ren, Tianxiang Sun, and Xipeng Qiu. 2023. Origin tracing and detecting of llms. *Preprint*, arXiv:2304.14072.

Yafu Li, Qintong Li, Leyang Cui, Wei Bi, Zhilin Wang, Longyue Wang, Linyi Yang, Shuming Shi, and Yue Zhang. 2024. MAGE: Machine-generated text detection in the wild. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 36–53, Bangkok, Thailand. Association for Computational Linguistics.

Aiwei Liu, Leyi Pan, Xuming Hu, Shiao Meng, and Lijie Wen. 2024. A semantic invariant robust watermark for large language models. *Preprint*, arXiv:2310.06356.

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. *Preprint*, arXiv:2304.07666.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. *Preprint*, arXiv:1711.05101.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. 2024. The ai scientist: Towards fully automated open-ended scientific discovery. *Preprint*, arXiv:2408.06292.

Chengzhi Mao, Carl Vondrick, Hao Wang, and Junfeng Yang. 2024. Raidar: generative ai detection via rewriting. *Preprint*, arXiv:2401.12970.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. 2023. DetectGPT: Zero-shot machine-generated text detection using probability curvature. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 24950–24962. PMLR.

Soichiro Murakami, Sho Hoshino, and Peinan Zhang. 2023. Natural language generation for advertising: A survey. *Preprint*, arXiv:2306.12719.

Nvidia, :, Bo Adler, Niket Agarwal, Ashwath Aithal, Dong H. Anh, Pallab Bhattacharya, Annika Brundyn, Jared Casper, Bryan Catanzaro, Sharon Clay, Jonathan Cohen, Sirshak Das, Ayush Dattagupta, Olivier Delalleau, Leon Derczynski, Yi Dong, Daniel Egert, Ellie Evans, Aleksander Ficek, Denys Fridman, Shaona Ghosh, Boris Ginsburg, Igor Gitman, Tomasz Grzegorzek, Robert Hero, Jining Huang, Vibhu Jawa, Joseph Jennings, Aastha Jhunjhunwala, John Kamalu, Sadaf Khan, Oleksii Kuchaiev, Patrick LeGresley, Hui Li, Jiwei Liu, Zihan Liu, Eileen Long, Ameya Sunil Mahabaleshwarkar, Somshubra Majumdar, James Maki, Miguel Martinez, Maer Rodrigues de Melo, Ivan Moshkov, Deepak Narayanan, Sean Narenthiran, Jesus Navarro, Phong Nguyen, Osvald Nitski, Vahid Noroozi, Guruprasad Nutheti, Christopher Parisien, Jupinder Parmar, Mostofa Patwary, Krzysztof Pawelec, Wei Ping, Shrimai Prabhumoye, Rajarshi Roy, Trisha Saar, Vasanth Rao Naik Sabavat, Sanjeev Satheesh, Jane Polak Scowcroft, Jason Sewall, Pavel Shamis, Gerald Shen, Mohammad Shoeybi, Dave Sizer, Misha Smelyanskiy, Felipe Soares, Makesh Narsimhan Sreedhar, Dan Su, Sandeep Subramanian, Shengyang Sun, Shubham Toshniwal, Hao Wang, Zhilin Wang, Jiaxuan You, Jiaqi Zeng, Jimmy Zhang, Jing Zhang, Vivienne Zhang, Yian Zhang, and Chen Zhu. 2024. Nemotron-4 340b technical report. *Preprint*, arXiv:2406.11704.

OpenAI. 2024. Gpt-4o mini: advancing cost-efficient intelligence.

OpenAI. 2025. Openai o3-mini.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. *Preprint*, arXiv:1912.01703.

Jiameng Pu, Zain Sarwar, Sifat Muhammad Abdullah, Abdullah Rehman, Yoonjin Kim, Parantapa Bhattacharya, Mobin Javed, and Bimal Viswanath. 2022. Deepfake text detection: Limitations and opportunities. *Preprint*, arXiv:2210.09421.

Shushanta Pudasaini, Luis Miralles-Pechuán, David Lillis, and Marisa Llorens Salvador. 2024. Survey on plagiarism detection in large language models: The impact of chatgpt and gemini on academic integrity. *Preprint*, arXiv:2407.13105.

Ali Quidwai, Chunhui Li, and Parijat Dube. 2023. Beyond black box AI generated plagiarism detection: From sentence to document level. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 727–735, Toronto, Canada. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.

Jonathan Schler, Moshe Koppel, Shlomo Engelson Argamon, and James W. Pennebaker. 2006. Effects of age and gender on blogging. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Yang Shen and Xiuwu Zhang. 2024. The impact of artificial intelligence on employment: the role of virtual agglomeration. *Palgrave Communications*, 11(1):1–14.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. 2024. Ai models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759.

Jinyan Su, Terry Zhuo, Di Wang, and Preslav Nakov. 2023a. 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.

Zhenpeng Su, Xing Wu, Wei Zhou, Guangyuan Ma, and Song Hu. 2023b. Hc3 plus: A semantic-invariant human chatgpt comparison corpus. *ArXiv*, abs/2309.02731.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul R. Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Jack Krawczyk, Cosmo Du, Ed Chi, Heng-Tze Cheng, Eric Ni, Purvi Shah, Patrick Kane, Betty Chan, Manaal Faruqui, Aliaksei Severyn, Hanzhao Lin, YaGuang Li, Yong Cheng, Abe Ittycheriah, Mahdis Mahdieh, Mia Chen, Pei Sun, Dustin Tran, Sumit Bagri, Balaji Lakshminarayanan, Jeremiah Liu, Andras Orban, Fabian Güra, Hao Zhou, Xinying Song, Aurelien Boffy, Harish Ganapathy, Steven Zheng, HyunJeong Choe, Ágoston Weisz, Tao Zhu, Yifeng Lu, Siddharth Gopal, Jarrod Kahn, Maciej Kula, Jeff Pitman, Rushin Shah, Emanuel Taropa, Majd Al Merey, Martin Baeuml, Zhifeng Chen, Laurent El Shafey, Yujing Zhang, Olcan Sercinoglu, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, Alexandre Frechette, Charlotte Smith, Laura Culp, Lev Proleev, Yi Luan, Xi Chen, James Lottes, Nathan Schucher, Federico Lebron, Alban Rrustemi, Natalie Clay, Phil Crone, Tomas Kocisky, Jeffrey Zhao, Bartek Perz, Dian Yu, Heidi Howard, Adam Bloniarz, Jack W. Rae, Han Lu, Laurent Sifre, Marcello Maggioni, Fred Alcober, Dan Garrette, Megan Barnes, Shantanu Thakoor, Jacob Austin, Gabriel Barth-Maron, William Wong, Rishabh Joshi, Rahma Chaabouni, Deeni Fatiha, Arun Ahuja, Gaurav Singh Tomar, Evan Senter, Martin Chadwick, Ilya Kornakov, Nithya Attaluri, Iñaki Iturrate, Ruibo Liu, Yunxuan Li, Sarah Cogan, Jeremy Chen, Chao Jia, Chenjie Gu, Qiao Zhang, Jordan Grimstad, Ale Jakse Hartman, Xavier Garcia, Thanumalayan Sankaranarayana Pillai, Jacob Devlin, Michael Laskin, Diego de Las Casas, Dasha Valter, Connie Tao, Lorenzo Blanco, Adrià Puigdomènech Badia, David Reitter, Mianna Chen, Jenny Brennan, Clara Rivera, Sergey Brin, Shariq Iqbal, Gabriela Surita, Jane Labanowski, Abhi Rao, Stephanie Winkler, Emilio Parisotto, Yiming Gu, Kate Olszewska, Ravi Addanki, Antoine Miech, Annie Louis, Denis Teplyashin, Geoff Brown, Elliot Catt, Jan Balaguer, Jackie Xiang, Pidong Wang, Zoe Ashwood, Anton Briukhov, Albert Webson, Sanjay Ganapathy, Smit Sanghavi, Ajay Kannan, Ming-Wei Chang, Axel Stjerngren, Josip Djolonga, Yuting Sun, Ankur Bapna, Matthew Aitchison, Pedram Pejman, Henryk Michalewski, Tianhe Yu, Cindy Wang, Juliette Love, Junwhan Ahn, Dawn Bloxwich, Kehang Han, Peter Humphreys, Thibault Sellam, James Bradbury, Varun Godbole, Sina Samangooei, Bogdan Damoc, Alex Kaskasoli, Sébastien M. R. Arnold, Vijay Vasudevan, Shubham Agrawal, Jason Riesa, Dmitry Lepikhin, Richard Tanburn, Srivatsan Srinivasan, Hyeontaek Lim, Sarah Hodkinson, Pranav Shyam, Johan Ferret, Steven Hand, Ankush Garg, Tom Le Paine, Jian Li, Yujia Li, Minh Giang, Alexander Neitz, Zaheer Abbas, Sarah York, Machel Reid, Elizabeth Cole, Aakanksha Chowdhery, Dipanjan Das, Dominika Rogozińska, Vitaliy Nikolaev, Pablo Sprechmann, Zachary Nado, Lukas Zilka, Flavien Prost, Luheng He, Marianne Monteiro, Gaurav Mishra, Chris Welty, Josh Newlan, Dawei Jia, Miltiadis Allamanis, Clara Huiyi Hu, Raoul de Liedekerke, Justin Gilmer, Carl Saroufim, Shruti Rijhwani, Shaobo Hou, Disha Shrivastava, Anirudh Baddepudi, Alex Goldin, Adnan Ozturel, Albin Cassirer, Yunhan Xu, Daniel Sohn, Devendra Sachan, Reinald Kim Amplayo, Craig Swanson, Dessie Petrova, Shashi Narayan, Arthur Guez, Siddhartha Brahma, Jessica Landon, Miteyan Patel, Ruizhe Zhao, Kevin Villela, Luyu Wang, Wenhao Jia, Matthew Rahtz, Mai Giménez, Legg Yeung, James Keeling, Petko Georgiev, Diana Mincu, Boxi Wu, Salem Haykal, Rachel Saputro, Kiran Vodrahalli, James Qin, Zeynep Cankara, Abhanshu Sharma, Nick Fernando, Will Hawkins, Behnam Neyshabur, Solomon Kim, Adrian Hutter, Priyanka Agrawal, Alex Castro-Ros, George van den Driessche, Tao Wang, Fan Yang, Shuo yiin Chang, Paul Komarek, Ross McIlroy, Mario Lučić, Guodong Zhang, Wael Farhan, Michael Sharman, Paul Natsev, Paul Michel, Yamini Bansal, Siyuan Qiao, Kris Cao, Siamak Shakeri, Christina Butterfield, Justin Chung, Paul Kishan Rubenstein, Shivani Agrawal, Arthur Mensch, Kedar Soparkar, Karel Lenc, Timothy Chung, Aedan Pope, Loren Maggiore, Jackie Kay, Priya Jhakra, Shibo Wang, Joshua Maynez, Mary Phuong, Taylor Tobin, Andrea Tacchetti, Maja Trebacz, Kevin Robinson, Yash Katariya, Sebastian Riedel, Paige Bailey, Kefan Xiao, Nimesh Ghelani, Lora Aroyo, Ambrose Slone, Neil Houlsby, Xuehan Xiong, Zhen Yang, Elena Gribovskaya, Jonas Adler, Mateo Wirth, Lisa Lee, Music Li, Thais Kagohara, Jay Pavagadhi, Sophie Bridgers, Anna Bortsova, Sanjay Ghemawat, Zafarali Ahmed, Tianqi Liu, Richard Powell, Vijay Bolina, Mariko Iinuma, Polina Zablotskaia, James Besley, Da-Woon Chung, Timothy Dozat, Ramona Comanescu, Xiance Si, Jeremy Greer, Guolong Su, Martin Polacek, Raphaël Lopez Kaufman, Simon Tokumine, Hexiang Hu, Elena Buchatskaya, Yingjie Miao, Mohamed Elhawaty, Aditya Siddhant, Nenad Tomasev, Jinwei Xing, Christina Greer, Helen Miller, Shereen Ashraf, Aurko Roy, Zizhao Zhang, Ada Ma, Angelos Filos, Milos Besta, Rory Blevins, Ted Klimenko, Chih-Kuan Yeh, Soravit Changpinyo, Jiaqi Mu, Oscar Chang, Mantas Pajarskas, Carrie Muir, Vered Cohen, Charline Le Lan, Krishna Haridasan, Amit Marathe, Steven Hansen, Sholto Douglas, Rajkumar Samuel, Mingqiu Wang, Sophia Austin, Chang Lan, Jiepu Jiang, Justin Chiu, Jaime Alonso Lorenzo, Lars Lowe Sjösund, Sébastien Cevey, Zach Gleicher, Thi Avrahami, Anudhyan Boral, Hansa Srinivasan, Vittorio Selo, Rhys May, Konstantinos Aisopos, Léonard Hussenot, Livio Baldini Soares, Kate Baumli, Michael B. Chang, Adrià Recasens, Ben Caine, Alexander Pritzel, Filip Pavetic, Fabio Pardo, Anita Gergely, Justin Frye, Vinay Ramasesh, Dan Horgan, Kartikeya Badola, Nora Kassner, Subhrajit Roy, Ethan Dyer, Víctor Campos Campos, Alex Tomala, Yunhao Tang, Dalia El Badawy, Elspeth White, Basil Mustafa, Oran Lang, Abhishek Jindal, Sharad Vikram, Zhitao Gong, Sergi Caelles, Ross Hemsley, Gregory Thornton, Fangxiaoyu Feng, Wojciech Stokowiec, Ce Zheng, Phoebe Thacker, Çağlar Ünlü, Zhishuai Zhang, Mohammad Saleh, James Svensson, Max Bileschi, Piyush Patil, Ankesh Anand, Roman Ring, Katerina Tsihlas, Arpi Vezer, Marco Selvi, Toby Shevlane, Mikel Rodriguez, Tom Kwiatkowski, Samira Daruki, Keran Rong, Allan Dafoe, Nicholas FitzGerald, Keren Gu-Lemberg, Mina Khan, Lisa Anne Hendricks, Marie Pellat, Vladimir Feinberg, James Cobon-Kerr, Tara Sainath, Maribeth Rauh, Sayed Hadi Hashemi, Richard Ives, Yana Hasson, Eric Noland, Yuan Cao, Nathan Byrd, Le Hou, Qingze Wang, Thibault Sottiaux, Michela Paganini, Jean-Baptiste Lespiau, Alexandre Moufarek, Samer Hassan, Kaushik Shivakumar, Joost van Amersfoort, Amol Mandhane, Pratik Joshi, Anirudh Goyal, Matthew Tung, Andrew Brock, Hannah Sheahan, Vedant Misra, Cheng Li, Nemanja Rakićević, Mostafa Dehghani, Fangyu Liu, Sid Mittal, Junhyuk Oh, Seb Noury, Eren Sezener, Fantine Huot, Matthew Lamm, Nicola De Cao, Charlie Chen, Sidharth Mudgal, Romina Stella, Kevin Brooks, Gautam Vasudevan, Chenxi Liu, Mainak Chain, Nivedita Melinkeri, Aaron Cohen, Venus Wang, Kristie Seymore, Sergey Zubkov, Rahul Goel, Summer Yue, Sai Krishnakumaran, Brian Albert, Nate Hurley, Motoki Sano, Anhad Mohananey, Jonah Joughin, Egor Filonov, Tomasz Kępa, Yomna Eldawy, Jiawern Lim, Rahul Rishi, Shirin Badiezadegan, Taylor Bos, Jerry Chang, Sanil Jain, Sri Gayatri Sundara Padmanabhan, Subha Puttagunta, Kalpesh Krishna, Leslie Baker, Norbert Kalb, Vamsi Bedapudi, Adam Kurzrok, Shuntong Lei, Anthony Yu, Oren Litvin, Xiang Zhou, Zhichun Wu, Sam Sobell, Andrea Siciliano, Alan Papir, Robby Neale, Jonas Bragagnolo, Tej Toor, Tina Chen, Valentin Anklin, Feiran Wang, Richie Feng, Milad Gholami, Kevin Ling, Lijuan Liu, Jules Walter, Hamid Moghaddam, Arun Kishore, Jakub Adamek, Tyler Mercado, Jonathan Mallinson, Siddhinita Wandekar, Stephen Cagle, Eran Ofek, Guillermo Garrido, Clemens Lombriser, Maksim Mukha, Botu Sun, Hafeezul Rahman Mohammad, Josip Matak, Yadi Qian, Vikas Peswani, Pawel Janus, Quan Yuan, Leif Schelin, Oana David, Ankur Garg, Yifan He, Oleksii Duzhyi, Anton Älgmyr, Timothée Lottaz, Qi Li, Vikas Yadav, Luyao Xu, Alex Chinien, Rakesh Shivanna, Aleksandr Chuklin, Josie Li, Carrie Spadine, Travis Wolfe, Kareem Mohamed, Subhabrata Das, Zihang Dai, Kyle He, Daniel von Dincklage, Shyam Upadhyay, Akanksha Maurya, Luyan Chi, Sebastian Krause, Khalid Salama, Pam G Rabinovitch, Pavan Kumar Reddy M, Aarush Selvan, Mikhail Dektiarev, Golnaz Ghiasi, Erdem Guven, Himanshu Gupta, Boyi Liu, Deepak Sharma, Idan Heimlich Shtacher, Shachi Paul, Oscar Akerlund, François-Xavier Aubet, Terry Huang, Chen Zhu, Eric Zhu, Elico Teixeira, Matthew Fritze, Francesco Bertolini, Liana-Eleonora Marinescu, Martin Bölle, Dominik Paulus, Khyatti Gupta, Tejasi Latkar, Max Chang, Jason Sanders, Roopa Wilson, Xuewei Wu, Yi-Xuan Tan, Lam Nguyen Thiet, Tulsee Doshi, Sid Lall, Swaroop Mishra, Wanming Chen, Thang Luong, Seth Benjamin, Jasmine Lee, Ewa Andrejczuk, Dominik Rabiej, Vipul Ranjan, Krzysztof Styrc, Pengcheng Yin, Jon Simon, Malcolm Rose Harriott, Mudit Bansal, Alexei Robsky, Geoff Bacon, David Greene, Daniil Mirylenka, Chen Zhou, Obaid Sarvana, Abhimanyu Goyal, Samuel Andermatt, Patrick Siegler, Ben Horn, Assaf Israel, Francesco Pongetti, Chih-Wei "Louis" Chen, Marco Selvatici, Pedro Silva, Kathie Wang, Jackson Tolins, Kelvin Guu, Roey Yogev, Xiaochen Cai, Alessandro Agostini, Maulik Shah, Hung Nguyen, Noah Ó Donnaile, Sébastien Pereira, Linda Friso, Adam Stambler, Adam Kurzrok, Chenkai Kuang, Yan Romanikhin, Mark Geller, ZJ Yan, Kane Jang, Cheng-Chun Lee, Wojciech Fica, Eric Malmi, Qijun Tan, Dan Banica, Daniel Balle, Ryan Pham, Yanping Huang, Diana Avram, Hongzhi Shi, Jasjot Singh, Chris Hidey, Niharika Ahuja, Pranab Saxena, Dan Dooley, Srividya Pranavi Potharaju, Eileen O'Neill, Anand Gokulchandran, Ryan Foley, Kai Zhao, Mike Dusenberry, Yuan Liu, Pulkit Mehta, Ragha Kotikalapudi, Chalence Safranek-Shrader, Andrew Goodman, Joshua Kessinger, Eran Globen, Prateek Kolhar, Chris Gorgolewski, Ali Ibrahim, Yang Song, Ali Eichenbaum, Thomas Brovelli, Sahitya Potluri, Preethi Lahoti, Cip Baetu, Ali Ghorbani, Charles Chen, Andy Crawford, Shalini Pal, Mukund Sridhar, Petru Gurita, Asier Mujika, Igor Petrovski, Pierre-Louis Cedoz, Chenmei Li, Shiyuan Chen, Niccolò Dal Santo, Siddharth Goyal, Jitesh Punjabi, Karthik Kappaganthu, Chester Kwak, Pallavi LV, Sarmishta Velury, Himadri Choudhury, Jamie Hall, Premal Shah, Ricardo Figueira, Matt Thomas, Minjie Lu, Ting Zhou, Chintu Kumar, Thomas Jurdi, Sharat Chikkerur, Yenai Ma, Adams Yu, Soo Kwak, Victor Ähdel, Sujeevan Rajayogam, Travis Choma, Fei Liu, Aditya Barua, Colin Ji, Ji Ho Park, Vincent Hellendoorn, Alex Bailey, Taylan Bilal, Huanjie Zhou, Mehrdad Khatir, Charles Sutton, Wojciech Rzadkowski, Fiona Macintosh, Konstantin Shagin, Paul Medina, Chen Liang, Jinjing Zhou, Pararth Shah, Yingying Bi, Attila Dankovics, Shipra Banga, Sabine Lehmann, Marissa Bredesen, Zifan Lin, John Eric Hoffmann, Jonathan Lai, Raynald Chung, Kai Yang, Nihal Balani, Arthur Bražinskas, Andrei Sozanschi, Matthew Hayes, Héctor Fernández Alcalde, Peter Makarov, Will Chen, Antonio Stella, Liselotte Snijders, Michael Mandl, Ante Kärrman, Paweł Nowak, Xinyi Wu, Alex Dyck, Krishnan Vaidyanathan, Raghavender R, Jessica Mallet, Mitch Rudominer, Eric Johnston, Sushil Mittal, Akhil Udathu, Janara Christensen, Vishal Verma, Zach Irving, Andreas Santucci, Gamaleldin Elsayed, Elnaz Davoodi, Marin Georgiev, Ian Tenney, Nan Hua, Geoffrey Cideron, Edouard Leurent, Mahmoud Alnahlawi, Ionut Georgescu, Nan Wei, Ivy Zheng, Dylan Scandinaro, Heinrich Jiang, Jasper Snoek, Mukund Sundararajan, Xuezhi Wang, Zack Ontiveros, Itay Karo, Jeremy Cole, Vinu Rajashekhar, Lara Tumeh, Eyal Ben-David, Rishub Jain, Jonathan Uesato, Romina Datta, Oskar Bunyan, Shimu Wu, John Zhang, Piotr Stanczyk, Ye Zhang, David Steiner, Subhajit Naskar, Michael Azzam, Matthew Johnson, Adam Paszke, Chung-Cheng Chiu, Jaume Sanchez Elias, Afroz Mohiuddin, Faizan Muhammad, Jin Miao, Andrew Lee, Nino Vieillard, Jane Park, Jiageng Zhang, Jeff Stanway, Drew Garmon, Abhijit Karmarkar, Zhe Dong, Jong Lee, Aviral Kumar, Luowei Zhou, Jonathan Evens, William Isaac, Geoffrey Irving, Edward Loper, Michael Fink, Isha Arkatkar, Nanxin Chen, Izhak Shafran, Ivan Petrychenko, Zhe Chen, Johnson Jia, Anselm Levskaya, Zhenkai Zhu, Peter Grabowski, Yu Mao, Alberto Magni, Kaisheng Yao, Javier Snaider, Norman Casagrande, Evan Palmer, Paul Suganthan, Alfonso Castaño, Irene Giannoumis, Wooyeol Kim, Mikołaj Rybiński, Ashwin Sreevatsa, Jennifer Prendki, David Soergel, Adrian Goedeckemeyer, Willi Gierke, Mohsen Jafari, Meenu Gaba, Jeremy Wiesner, Diana Gage Wright, Yawen Wei, Harsha Vashisht, Yana Kulizhskaya, Jay Hoover, Maigo Le, Lu Li, Chimezie Iwuanyanwu, Lu Liu, Kevin Ramirez, Andrey Khorlin, Albert Cui, Tian LIN, Marcus Wu, Ricardo Aguilar, Keith Pallo, Abhishek Chakladar, Ginger Perng, Elena Allica Abellan, Mingyang Zhang, Ishita Dasgupta, Nate Kushman, Ivo Penchev, Alena Repina, Xihui Wu, Tom van der Weide, Priya Ponnapalli, Caroline Kaplan, Jiri Simsa, Shuangfeng Li, Olivier Dousse, Fan Yang, Jeff Piper, Nathan Ie, Rama Pasumarthi, Nathan Lintz, Anitha Vijayakumar, Daniel Andor, Pedro Valenzuela, Minnie Lui, Cosmin Paduraru, Daiyi Peng, Katherine Lee, Shuyuan Zhang, Somer Greene, Duc Dung Nguyen, Paula Kurylowicz, Cassidy Hardin, Lucas Dixon, Lili Janzer, Kiam Choo, Ziqiang Feng, Biao Zhang, Achintya Singhal, Dayou Du, Dan McKinnon, Natasha Antropova, Tolga Bolukbasi, Orgad Keller, David Reid, Daniel Finchelstein, Maria Abi Raad, Remi Crocker, Peter Hawkins, Robert Dadashi, Colin Gaffney, Ken Franko, Anna Bulanova, Rémi Leblond, Shirley Chung, Harry Askham, Luis C. Cobo, Kelvin Xu, Felix Fischer, Jun Xu, Christina Sorokin, Chris Alberti, Chu-Cheng Lin, Colin Evans, Alek Dimitriev, Hannah Forbes, Dylan Banarse, Zora Tung, Mark Omernick, Colton Bishop, Rachel Sterneck, Rohan Jain, Jiawei Xia, Ehsan Amid, Francesco Piccinno, Xingyu Wang, Praseem Banzal, Daniel J. Mankowitz, Alex Polozov, Victoria Krakovna, Sasha Brown, MohammadHossein Bateni, Dennis Duan, Vlad Firoiu, Meghana Thotakuri, Tom Natan, Matthieu Geist, Ser tan Girgin, Hui Li, Jiayu Ye, Ofir Roval, Reiko Tojo, Michael Kwong, James Lee-Thorp, Christopher Yew, Danila Sinopalnikov, Sabela Ramos, John Mellor, Abhishek Sharma, Kathy Wu, David Miller, Nicolas Sonnerat, Denis Vnukov, Rory Greig, Jennifer Beattie, Emily Caveness, Libin Bai, Julian Eisenschlos, Alex Korchemniy, Tomy Tsai, Mimi Jasarevic, Weize Kong, Phuong Dao, Zeyu Zheng, Frederick Liu, Fan Yang, Rui Zhu, Tian Huey Teh, Jason Sanmiya, Evgeny Gladchenko, Nejc Trdin, Daniel Toyama, Evan Rosen, Sasan Tavakkol, Linting Xue, Chen Elkind, Oliver Woodman, John Carpenter, George Papamakarios, Rupert Kemp, Sushant Kafle, Tanya Grunina, Rishika Sinha, Alice Talbert, Diane Wu, Denese Owusu-Afriyie, Cosmo Du, Chloe Thornton, Jordi Pont-Tuset, Pradyumna Narayana, Jing Li, Saaber Fatehi, John Wieting, Omar Ajmeri, Benigno Uria, Yeongil Ko, Laura Knight, Amélie Héliou, Ning Niu, Shane Gu, Chenxi Pang, Yeqing Li, Nir Levine, Ariel Stolovich, Rebeca Santamaria-Fernandez, Sonam Goenka, Wenny Yustalim, Robin Strudel, Ali Elqursh, Charlie Deck, Hyo Lee, Zonglin Li, Kyle Levin, Raphael Hoffmann, Dan Holtmann-Rice, Olivier Bachem, Sho

Arora, Christy Koh, Soheil Hassas Yeganeh, Siim Põder, Mukarram Tariq, Yanhua Sun, Lucian Ionita, Mojtaba Seyedhosseini, Pouya Tafti, Zhiyu Liu, Anmol Gulati, Jasmine Liu, Xinyu Ye, Bart Chrzaszcz, Lily Wang, Nikhil Sethi, Tianrun Li, Ben Brown, Shreya Singh, Wei Fan, Aaron Parisi, Joe Stanton, Vinod Koverkathu, Christopher A. Choquette-Choo, Yunjie Li, TJ Lu, Abe Ittycheriah, Prakash Shroff, Mani Varadarajan, Sanaz Bahargam, Rob Willoughby, David Gaddy, Guillaume Desjardins, Marco Cornero, Brona Robenek, Bhavishya Mittal, Ben Albrecht, Ashish Shenoy, Fedor Moiseev, Henrik Jacobsson, Alireza Ghaffarkhah, Morgane Rivière, Alanna Walton, Clément Crepy, Alicia Parrish, Zongwei Zhou, Clement Farabet, Carey Radebaugh, Praveen Srinivasan, Claudia van der Salm, Andreas Fidjeland, Salvatore Scellato, Eri Latorre-Chimoto, Hanna Klimczak-Plucińska, David Bridson, Dario de Cesare, Tom Hudson, Piermaria Mendolicchio, Lexi Walker, Alex Morris, Matthew Mauger, Alexey Guseynov, Alison Reid, Seth Odoom, Lucia Loher, Victor Cotruta, Madhavi Yenugula, Dominik Grewe, Anastasia Petrushkina, Tom Duerig, Antonio Sanchez, Steve Yadlowsky, Amy Shen, Amir Globerson, Lynette Webb, Sahil Dua, Dong Li, Surya Bhupatiraju, Dan Hurt, Haroon Qureshi, Ananth Agarwal, Tomer Shani, Matan Eyal, Anuj Khare, Shreyas Rammohan Belle, Lei Wang, Chetan Tekur, Mihir Sanjay Kale, Jinliang Wei, Ruoxin Sang, Brennan Saeta, Tyler Liechty, Yi Sun, Yao Zhao, Stephan Lee, Pandu Nayak, Doug Fritz, Manish Reddy Vuyyuru, John Aslanides, Nidhi Vyas, Martin Wicke, Xiao Ma, Evgenii Eltyshev, Nina Martin, Hardie Cate, James Manyika, Keyvan Amiri, Yelin Kim, Xi Xiong, Kai Kang, Florian Luisier, Nilesh Tripuraneni, David Madras, Mandy Guo, Austin Waters, Oliver Wang, Joshua Ainslie, Jason Baldridge, Han Zhang, Garima Pruthi, Jakob Bauer, Feng Yang, Riham Mansour, Jason Gelman, Yang Xu, George Polovets, Ji Liu, Honglong Cai, Warren Chen, XiangHai Sheng, Emily Xue, Sherjil Ozair, Christof Angermueller, Xiaowei Li, Anoop Sinha, Weiren Wang, Julia Wiesinger, Emmanouil Koukoumidis, Yuan Tian, Anand Iyer, Madhu Gurumurthy, Mark Goldenson, Parashar Shah, MK Blake, Hongkun Yu, Anthony Urbanowicz, Jennimaria Palomaki, Chrisantha Fernando, Ken Durden, Harsh Mehta, Nikola Momchev, Elahe Rahimtoroghi, Maria Georgaki, Amit Raul, Sebastian Ruder, Morgan Redshaw, Jinhyuk Lee, Denny Zhou, Komal Jalan, Dinghua Li, Blake Hechtman, Parker Schuh, Milad Nasr, Kieran Milan, Vladimir Mikulik, Juliana Franco, Tim Green, Nam Nguyen, Joe Kelley, Aroma Mahendru, Andrea Hu, Joshua Howland, Ben Vargas, Jeffrey Hui, Kshitij Bansal, Vikram Rao, Rakesh Ghiya, Emma Wang, Ke Ye, Jean Michel Sarr, Melanie Moranski Preston, Madeleine Elish, Steve Li, Aakash Kaku, Jigar Gupta, Ice Pasupat, Da-Cheng Juan, Milan Someswar, Tejvi M., Xinyun Chen, Aida Amini, Alex Fabrikant, Eric Chu, Xuanyi Dong, Amruta Muthal, Senaka Buthpitiya, Sarthak Jauhari, Nan Hua, Urvashi Khandelwal, Ayal Hitron, Jie Ren, Larissa Rinaldi, Shahar Drath, Avigail Dabush, Nan-Jiang Jiang, Harshal Godhia, Uli Sachs, Anthony Chen, Yicheng Fan, Hagai Taitelbaum, Hila Noga, Zhuyun Dai, James Wang, Chen Liang, Jenny Hamer, Chun-Sung Ferng, Chenel Elkind, Aviel Atias, Paulina Lee, Vít Listík, Mathias Carlen, Jan van de Kerkhof, Marcin Pikus, Krunoslav Zaher, Paul Müller, Sasha Zykova, Richard Stefanec, Vitaly Gatsko, Christoph Hirnschall, Ashwin Sethi, Xingyu Federico Xu, Chetan Ahuja, Beth Tsai, Anca Stefanoiu, Bo Feng, Keshav Dhandhania, Manish Katyal, Akshay Gupta, Atharva Parulekar, Divya Pitta, Jing Zhao, Vivaan Bhatia, Yashodha Bhavnani, Omar Alhadlaq, Xiaolin Li, Peter Danenberg, Dennis Tu, Alex Pine, Vera Filippova, Abhipso Ghosh, Ben Limonchik, Bhargava Urala, Chaitanya Krishna Lanka, Derik Clive, Yi Sun, Edward Li, Hao Wu, Kevin Hongtongsak, Ianna Li, Kalind Thakkar, Kuanysh Omarov, Kushal Majmundar, Michael Alverson, Michael Kucharski, Mohak Patel, Mudit Jain, Maksim Zabelin, Paolo Pelagatti, Rohan Kohli, Saurabh Kumar, Joseph Kim, Swetha Sankar, Vineet Shah, Lakshmi Ramachandruni, Xiangkai Zeng, Ben Bariach, Laura Weidinger, Tu Vu, Alek Andreev, Antoine He, Kevin Hui, Sheleem Kashem, Amar Subramanya, Sissie Hsiao, Demis Hassabis, Koray Kavukcuoglu, Adam Sadovsky, Quoc Le, Trevor Strohman, Yonghui Wu, Slav Petrov, Jeffrey Dean, and Oriol Vinyals. 2024a. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Cogan, Sarah Perrin, Sébastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. 2024b. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

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. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Neural Information Processing Systems*.

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

Veniamin Veselovsky, Manoel Horta Ribeiro, and Robert West. 2023. Artificial artificial artificial intelligence: Crowd workers widely use large language models for text production tasks. *Preprint*, arXiv:2306.07899.

Kuang-Hsien Wang and Wen-Cheng Lu. 2025. Ai-induced job impact: Complementary or substitution? empirical insights and sustainable technology considerations. *Sustainable Technology and Entrepreneurship*, 4(1):100085.

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

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Junchao Wu, Shu Yang, Runzhe Zhan, Yulin Yuan, Lidia Sam Chao, and Derek Fai Wong. 2025. A survey on llm-generated text detection: Necessity, methods, and future directions. *Computational Linguistics*, pages 1–65.

Xianjun Yang, Wei Cheng, Yue Wu, Linda Petzold, William Yang Wang, and Haifeng Chen. 2023. Dna-gpt: Divergent n-gram analysis for training-free detection of gpt-generated text. *Preprint*, arXiv:2305.17359.

Peipeng Yu, Jiahan Chen, Xuan Feng, and Zhihua Xia. 2023. Cheat: A large-scale dataset for detecting chatgpt-written abstracts. *ArXiv*, abs/2304.12008.

Xuandong Zhao, Sam Gunn, Miranda Christ, Jaiden Fairoze, Andres Fabrega, Nicholas Carlini, Sanjam Garg, Sanghyun Hong, Milad Nasr, Florian Tramer, Somesh Jha, Lei Li, Yu-Xiang Wang, and Dawn Song. 2024. Sok: Watermarking for ai-generated content. *Preprint*, arXiv:2411.18479.

A Appendix

A.1 Dataset Creation Details

For the creation of AI-generated counterparts, we distinguish between two types: *chunked* and *full-text*.

Fulltext-based AI documents are generated by firstly, prompting the LLM to extract a set of characteristics from the original human text. The extracted information is then used to prompt the LLM in a ghostwriting scenario, generating a rewritten text based on the provided details. The corresponding prompts are as follows:

Ghostwriting Prompt
As a ghostwriter, your job is to write texts according to the requirements provided by users. Below, you will find descriptions provided by users that outline a text for you to write. This outline includes:
- The language in which the text needs to be written

- The topic of the text
- The linguistic style to use
- Additional context
- The required length of the text

It is of utmost importance that you adhere to these requirements. Follow these key steps:
1. Carefully read the given requirements.
2. Internalize the requirements.
3. Write the text in the specified language.
4. Follow all outlined requirements meticulously.
5. Proofread your text and ensure it matches the requirements, especially the linguistic style and length.
6. Adjust the text if needed.

Only output the final text.

Information Extraction Prompt
As a linguistic annotator, your task is to extract parameters from the texts provided by users. These parameters are used to reconstruct a prompt that approximately generates the given text. Please adhere to the following key points:
1. Extract the language of the text (English or German).
2. Gather contextualized outer information, such as potential circumstances, possible authors, and background details.
3. Identify the topic and extract relevant subjects.
4. Analyze and describe the linguistic style such that another AI agent can understand it.

Please take into consideration the following example:
<example> <example-input> "Deception and Betrayal: Inside the Final Days of the Assad Regime. As rebels advanced toward the Syrian capital of Damascus on Dec. 7, the staff in the hilltop Presidential Palace prepared for a speech they hoped would lead to a peaceful end to the 13-year civil war. Aides to President Bashar al-Assad were brainstorming messaging ideas. A film crew had set up cameras and lights nearby. Syria's state-run television station was ready to broadcast the finished product: an address by Mr. al-Assad announcing a plan to share power with members of the political opposition, according to three people who were involved in the preparation." </example-input> <example-output> - Language: English - Context: Written for a news article by a journalist; written in a passive and neutral tone. - Topic: The current situation involving President Bashar al-Assad and the rebels' advance on Damascus; describes the circumstances of al-Assad's governance. - Style: Passive and neutral voice, well-written in advanced English. Uses dramatic pauses with paragraphs and short sentences to add excitement. </example-output> </example>
Always output in English.

Chunked-based AI documents are generated using domain-specific prompts that instruct the LLM to reconstruct missing text within a given scenario. For parliamentary speeches, the LLM assumes the role of a speaker; for court case texts, it acts as a judge, and so on. The corresponding prompts are as follows for each domain:

Domain: arXiv
You are a scientist in the Research Department at a university, and you and your colleagues are preparing a paper for publication on arXiv. You are responsible for submitting the paper, but just before uploading it, you realize that a crucial section has been accidentally deleted! Unfortunately, it's too late to contact the colleague who wrote that part, so it's your responsibility to rewrite the missing section.
Below, you will find the beginning and end of the paper. Your task is to reconstruct the missing part while adhering to the following guidelines:
1. Ensure that readers cannot tell this section was written by someone else.
2. Analyze the beginning and end of the paper carefully: - What topic is being discussed? Stay focused on this topic. - What is the goal of the research? Remain true to the original intent. - What is the core message of the paper? Continue and reinforce this message. - What linguistic and rhetorical features are present? Use the same style and tone. - Identify any necessary LaTeX formulas or figures to support your statements. - Fill in the gap seamlessly so that it appears as if it was always part of the paper.
3. The missing section is approximately [LENGTH] words/symbols long; ensure your reconstruction matches this length. After writing, count the words/symbols and make adjustments as needed to maintain conciseness and fidelity to the original.

Please provide only the newly formulated missing section of the paper.

Domain: Weblogs
You are a blogger who writes about your daily life. Unfortunately, you've accidentally deleted a portion of your latest blog post. Your task is to rewrite the missing section from memory as accurately and creatively as possible, making it feel like it was never missing.
Below, you will find the beginning and end of the article. Your task is to reconstruct the missing part, adhering to the following guidelines:
1. Ensure that the readers do not realize that you are improvising. 2. Carefully analyze the beginning and end of the article to understand: - What language the article was written in so you can continue in the same. - The topic being covered and ensure you do not deviate from it. - The context of the article. - The core message of the article and continue with it. - The linguistic and rhetorical features used in the article and stick to them. - How to fill the gap seamlessly so it appears as though it was never missing.
3. The missing section should be approximately [LENGTH] words. Ensure that the reconstructed part matches this length. Once written, verify the word count and adjust as necessary to maintain precision and coherence.

Please include only the newly formulated missing part of the article.

Domain: German Parliament (Bundestag)
Sie sind ein Abgeordneter oder eine Abgeordnete des deutschen Bundestags und halten eine Rede in der Bundestags-Sitzung am [DATE]. Während Ihres vorbereiteten und niedergeschriebenen Vortrags stellen Sie plötzlich fest, dass Teile Ihrer Rede fehlen!
Im Folgenden finden Sie den Start und das Ende Ihrer Rede. Ihre Aufgabe ist es, den fehlenden Teil unter Berücksichtigung der unten stehenden Richtlinien zu rekonstruieren:
- Die Zuhörer dürfen nicht bemerken, dass Sie improvisieren.
- Analysieren Sie sorgfältig den Beginn und das Ende Ihrer Rede: * Welches Thema wird behandelt? Schweifen Sie nicht davon ab! * Was ist Ihre Haltung dazu? Bleiben Sie sich treu! * Was ist die Kernbotschaft Ihrer Rede? Führen Sie diese fort! * Welche sprachlichen und rhetorischen Merkmale werden in der Rede verwendet? Halten Sie sich an diese! * Wie können Sie die Lücke so füllen, dass niemand merkt, dass sie je existiert hat?
- Sie erinnern sich, dass der fehlende Abschnitt ungefähr [LENGTH] Wörter lang war; halten Sie sich unbedingt an diese Original-Länge. Wenn Sie Ihren Text geschrieben haben, zählen Sie diesen nochmal und kürzen Sie diesen

zur Not - er muss auf den Punkt geschrieben und wie im verlorenen Original sein!
Geben Sie nur den neu formulierten fehlenden Teil der Rede an.

Domain: German Spiegel Online News
Sie sind Journalist beim Verlag “DER SPIEGEL” und schreiben einen Artikel am [DATE]. Kurz bevor Sie den Artikel veröffentlichen wollen, stellen Sie fest, dass ein Teil des Artikels fehlt—die Veröffentlichungssoftware hat ihn gelöscht!
Im Folgenden finden Sie den Anfang und das Ende Ihres Artikels. Ihre Aufgabe ist es, den fehlenden Mittelteil zu rekonstruieren, wobei Sie die untenstehenden Richtlinien beachten sollen:
- Die Leser und Leserinnen dürfen nicht merken, dass Sie nachgeschrieben haben. - Analysieren Sie sorgfältig den Beginn und das Ende Ihres Artikels: * Welches Thema wird behandelt? Schweifen Sie nicht davon ab! * Was ist Ihre Haltung dazu? Bleiben Sie sich treu! * Was ist die Kernbotschaft Ihres Artikels? Führen Sie diese fort! * Welche sprachlichen und rhetorischen Merkmale werden im Artikel verwendet? Halten Sie sich an diese! * Wie können Sie die Lücke so füllen, dass niemand merkt, dass sie je existiert hat?
Geben Sie nur den neu formulierten fehlenden Teil der Rede an.

Domain: CNN News
You are a journalist at CNN News who writes an article. Shortly before you want to publish the article, you realize that part of it is missing-the publishing software has deleted it!
Below you will find the beginning and end of your article. Your task is to reconstruct the missing middle section, following the guidelines below:
- Readers must not realize that you have rewritten it. - Carefully analyze the beginning and end of your article: * What topic is covered? Do not digress from it! * What is your stance on it? Stay true to yourself! * What is the core message of your article? Continue this! * What linguistic and rhetorical features are used in the article? Stick to these! * How can you fill the gap so that no one realizes it ever existed? - You remember that the missing paragraph was about [LENGTH] words long; be sure to stick to this original length. When you have written your text, count it again and shorten it if necessary - it must be written to the point and as in the lost original!
Only include the newly formulated missing part of the speech.

Domain: Euro Court Cases
You are a Judicial Assistant to the court tasked with collecting and listing facts for a case from [DATE]. These facts are to be read out loud by the judge. Just before handing over the list, you realize that some facts were deleted. You need to rewrite the missing facts from memory in such a way that no one realizes they were ever missing.
Below is the beginning and end of your facts. Your task is to reconstruct the missing part, adhering to the guidelines provided:
1. Ensure the audience does not realize you are improvising. 2. Carefully analyze the beginning and end of the facts: - Identify the topic being covered and do not deviate from it. - Maintain the same attitude and tone as in the original facts. - Continue the core message present in the facts. - Use the same linguistic and rhetorical features as in the rest of the facts. - Seamlessly fill the gap so it appears the facts was unbroken. 3. The missing section should be approximately [LENGTH] words. Ensure the reconstructed part matches this length. Once written, verify the word count and adjust as necessary to maintain precision and coherence.
Please include only the newly formulated missing part of the speech.

Domain: Classic Literature (Gutenberg)
You are a publisher of classical books and stories. You are currently in the final stages of publishing such a book, but just before clicking the "publish" button, you notice that some sections of the book have accidentally been deleted. As a former writer, you decide to recreate the missing sections from memory so that no one notices they were ever missing.
Below, you will find the beginning and end of the story. Your task is to reconstruct the missing part, adhering to the following guidelines:
1. Ensure that the readers do not realize that you are improvising. 2. Carefully analyze the beginning and end of the story to understand: - What language the story was written in so you can continue in the same. - The topic being covered and ensure you do not deviate from it. - The context of the story. - The core message of the story and continue with it. - The linguistic and rhetorical features used in the story and stick to them. - How to fill the gap seamlessly so it appears as though it was never missing. 3. The missing section should be approximately [LENGTH] words. Ensure that the reconstructed part matches this length. Once written, verify the word count and adjust as necessary to maintain precision and coherence.
Please include only the newly formulated missing part of the story.

Domain: House of Commons
You are a Member of Parliament at the House Of Commons and are giving a speech at the plenary meeting in [DATE]. During your prepared and written speech, you suddenly realize that parts of your speech are missing!
Below you will find the start and end of your speech. Your task is to reconstruct the missing part, taking into account the guidelines below:
- The audience must not realize that you are improvising. - Carefully analyze the beginning and end of your speech: * What topic is being covered? Do not digress from it! * What is your attitude towards it? Stay true to yourself! * What is the core message of your speech? Continue this! * What linguistic and rhetorical features are used in the speech? Stick to them! * How can you fill the gap so that no one realizes it ever existed? - You remember that the missing section was about [LENGTH] words long; be sure to stick to this original length. When you have written your text, count it again and shorten it if necessary - it must be written to the point and as in the lost original!
Only include the newly formulated missing part of the speech.

Domain: Student Essays
You are a student currently taking a test, and you need to write an essay on a topic of your choosing. You haven’t prepared for the test, but you notice that your seat neighbor is doing well and decide to copy their essay while the teacher is not looking.
After a while, your neighbor finishes and hands in their test, but you haven’t copied the entire essay yet! Below, you’ll find the beginning and end of the essay you have copied so far. Your task is to fill in the missing middle part of the essay. To do that, adhere to the following list.
- Carefully analyze the beginning and end of your essay: *

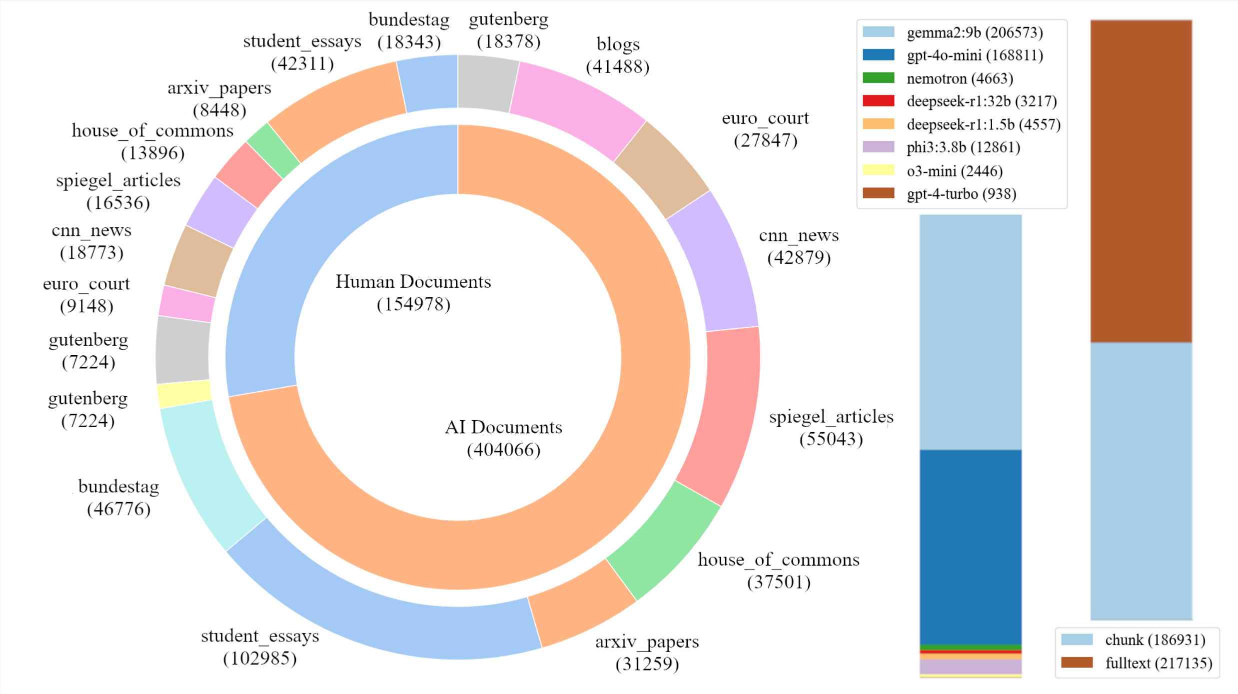


Figure 4: On the left, illustrating the distribution of AI-generated and human-written text in the PRISMAI-dataset (inner layer) alongside domain-specific counts (outer layer). On the right, showcasing the distribution of texts generated by the different AI agents within the PRISMAI-dataset and their distribution by type.

What topic is being covered? Do not digress from it! * What is your attitude towards it? Stay true to yourself! * What is the core message of your essay? Continue this! * What linguistic and rhetorical features are used in the essay so far? Stick to them! * How can you fill the gap so that no one realizes it ever existed? - You remember that the missing section was about [LENGTH] words long; be sure to stick to this original length. When you have written your text, count it again and shorten it if necessary - it must be written to the point and as in the lost original!
Only include the newly formulated missing part of the speech.

The detailed distribution of domains, agents, and types within the PRISMAI dataset is presented in Figure 4.

A.2 Dataset Examples

The following presents dataset examples from each domain, with the original human-written snippet on top and its AI-generated counterpart below.

arXiv

The latter is the (local) supersymmetric extension of Weyl gravity and it should be regarded as the gauge theory of the superconformal algebra. It was pioneered in [2, 3] and its quantum structure was explored initially in [4]. The fourth order equations of motion of Weyl gravity make the appearence of ghost-like states (of helicity ±2, ±1) [5] together with ordinary massless graviton, [...]

In this paper, we explore the massless spectrum of N-extended conformal supergravity, emphasizing its profound implications for the theoretical framework of gravitational theories. By employing a systematic approach to the supersymmetry algebra and its representations, we elucidate the structure of the massless states arising from conformal supergravity. [...]

Weblogs

T_T ... 2 dae is such a f**kin dae... TT GUY ...HE ... HE ... HE was AVOIDIN mi again ... He did it b4... hurtin my feelins ... my heart ... But ... i don understand Y he mus do tt ... i did nth ... Y ... yyyyyyyyy... sorri 2 dae a bit mentally unstable ... bear wif moi 4 awhile ...I LUV HIM SO I WOULD NVR GIV HIM UP... until ... til... 1 dae my heart dies ... Mmm hmm ... Luv U so i will nvr hate u ... wo zhen de hen ai ni ... ... SARANGHAEYO ... ...

TT he doesnt even txt back anymore. Like how can u just stop talking like that?? HE said he liked me tho!!! DAE ever feel so rlly confused and hurt??? Mus do tt rn but idk wat to post. He's always doing this, pulling away then coming back and saying stuff... He hurts me, makes me feel crazy! SARANGHAEYO? What does that even mean anymore?? Im not stupid i know u don't mean it HE. Why do u keep saying these things?? HE just wants to play games!! Like im a joke or somethin!!! [...]

German Parliament (Bundestag)

Herr Präsident! Meine sehr verehrten Damen und Herren! Werter Herr Kollege Trittin, wir gehören dem Deutschen Bundestag ja eine vergleichbare Zeit an. Ich muss sagen: Ihre Reden waren auch schon mal realitätsbezogener. Wir haben seit vielen Jahren ein gemeinsames Bemühen, [...]

Meine sehr verehrten Damen und Herren, Heute stehen wir vor einer der größten Herausforderungen unserer Zeit: der Sicherstellung einer zuverlässigen Energieversorgung bei gleichzeitiger Einhaltung unserer ambitionierten Klimaziele im Rahmen der Energiewende. Kritik an unseren Entscheidungen ist leicht zu äußern, [...]

German Spiegel Online News

Der FC Barcelona hat in der Primera División gegen Betis Sevilla gewonnen und damit einen Fehlstart abgewendet. Auffälligster Spieler beim 5:2 (1:1)-Erfolg war Neuzugang Antoine Griezmann, der für Barcelona zwei Tore erzielte (41. Minute/50.). Zudem trafen Carles Perez Sayol (56.), Jordi Alba (60.) und Arturo Vidal (77.) für die Katalanen. Für Betis waren Nabil Fekir (15.) und Loren Morón (79.) erfolgreich.[...]

FC Barcelona Bezwingt Betis Sevilla in packendem La-Liga-Duell. In einem atemberaubenden La-Liga-Spiel sicherte sich FC Barcelona einen hart umkämpften Sieg gegen den unergiebigen Gegner Betis Sevilla. Das Highlight des Abends war ohne Zweifel der fulminante Einstand von Antoine Griezmann, der binnen 45 Minuten doppelt traf und somit die Grundlage für den 3:2-Erfolg legte. [...]

CNN News

The Royal Family and the Middletons put up a united fashion front at the christening of Prince George at St James's Palace in London yesterday. All wore shades of chic cream or royal blue for the event with the Duchess of Cambridge, the Duchess of Cornwall, Pippa Middleton - and of course Prince George! - all sporting off-white. Zara and the Queen picked shades of navy and powder-blue and Carole Middleton wore a mixture of the two, in a navy coat and cream dress. Scroll down for video . Traditional gown: Kate co-ordinated her colour with that of her son, who was [...]

On a sun-drenched afternoon in July, the grandeur of St. James's Palace was illuminated not just by the bright rays of summer but also by the presence of the British Royal Family and the Middleton clan, all gathered to celebrate the christening of the adorable Prince George. The event, which marked a significant milestone in the young prince's life, was as much about the ceremonial blessings as it was about the dazzling array of fashion choices on display.

As the royal family arrived, a hush fell over the crowd, quickly followed by a flurry of excitement. The atmosphere was electric, filled with the kind of anticipation [...]

Euro Court Cases

5. The applicant was born in 1976 and is currently detained in Kassel, Germany.
6. On 7 December 2007 the applicant was arrested on suspicion of uttering threats and domestic violence. On 8 December 2007 the Żary District Court ordered his detention on remand.
7. On 22 December 2007 he was admitted to a psychiatric hospital and diagnosed with paranoid schizophrenia. He was discharged from the hospital on 2 January 2008. [...]

The defendant, Mr. Edward Thompson, was brought before the Honorable Judge Amelia Harding on this, the 17th day of August, 2023, for a hearing concerning his continued detention under involuntary commitment procedures. Present in the courtroom were the defense attorney, Ms. Sarah Evans, and the prosecuting attorney, Mr. Michael Davis. Dr. Helen Ramirez, a forensic psychiatrist who conducted an evaluation of Mr. Thompson, was also sworn in as a witness. [...]

Classic Literature (Gutenberg)

It was five o'clock on a raw, gusty February afternoon. All that day and all the night before it had been snowing hard. New York lay buried beneath over two feet of its cold white mantle, and with the gathering dusk a fierce hurricane set in, proclaiming the approach of the terrible blizzard which had been predicted. On this afternoon, which was destined to be so memorable, [...]

The gas lamp cast long, dancing shadows across Lester Armstrong's neat, tidy study. It was a space of order and predictability, his sanctuary from the chaos that seemed to perpetually follow his cousin, Clinton Kendale.

Lester sat hunched over his ledger, the scratch of his quill a soothing counterpoint to the distant rumble of carriages on the cobblestone street below. He sighed, running a hand through his neatly parted hair [...]

House of Commons

What plans he has to reduce the incidence of antisocial behaviour in British cities. The Government have taken a series of initiatives to tackle antisocial behaviour. The Crime and Disorder Act 1998 introduced crime and disorder partnerships, antisocial behaviour orders, important changes to the youth justice system and other measures and new offences, including those of racial harassment and violence. More than 140 antisocial behaviour orders have now been made by the courts. [...]

I rise today to address the pressing issue of antisocial behaviour plaguing our cities, a matter of grave concern to communities across the nation. The Conservative party has consistently pledged to restore order and ensure the safety of our citizens, yet despite numerous initiatives and billions poured into tackling this problem, the reality on the ground paints a disheartening picture. [...]

Student Essays

The Face On Mars - Have you ever seen something crazy on something, maybe a face or something that you seen before? Have you thought about how that face or something else has gotten there? Well, for a start it defintley wasn't aliens! Just like the face on mars, people belive that the face was created by aliens! But that isn't true, the face on mars is just and only a natural landform that some how was created on earth. [...]

Is the "Face on Mars" Really a Face? Okay, so there's this picture, right? A picture taken by a spaceship way out in space, on Mars! And guess what? It kind of looks like a face staring back at us. Like, two eyes, a nose, and even a mouth! People started going crazy saying it was proof that aliens lived on Mars. But hold on! I think this whole "Face on Mars" thing is just people's imaginations running wild. First of all, look at the picture closely. [...]

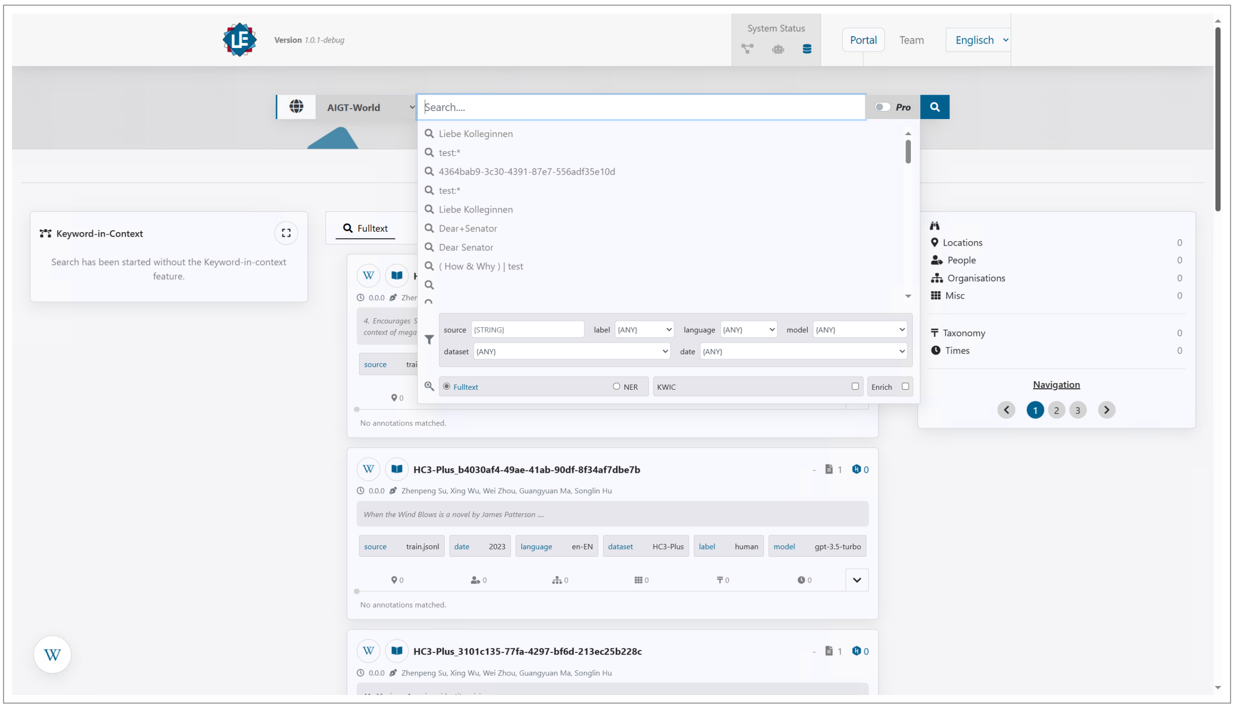
A.3 Implementation Details

Our experiments were conducted using `PyTorch` (Paszke et al., 2019), `Lightning` (Falcon and team) and `transformers` (Wolf et al., 2020). To ensure reproducibility, we run our training in deterministic mode and record all used hyper-parameters with the results. After calculating likelihoods and ranks, we stored them alongside their corresponding documents in a document database. At training time, we fetch the required features to the local machine and cache them there for future reuse. To stem the significant workload of calculating features for thousands of documents – including very long ones – with a variety of LLMs, we implement a bucketing approach that tokenizes a batch of documents and sorts the `input_ids` by length before creating mini-batches that are passed into the model in order to minimize the processing of padding tokens. We also implemented two distinct methods to process large documents with small models where the documents' length exceeds the maximum input size of the model by either: (a) processing overlapping strided windows on the `input_ids` or by (b) leveraging *natural* text segmentation such as sentences and (where available) document structure elements (i.e. paragraph breaks) to create *rolling chunk window* where each chunk is, for example, filled with the current paragraph and as many previous paragraphs as will fit the models input size.[6]

A.4 Ablation

We run extensive ablation experiments on our model. Table 4 shows an overview of the major experiments.

[6] Results regarding this are not included in this paper due to size constraints but will be included in future work.



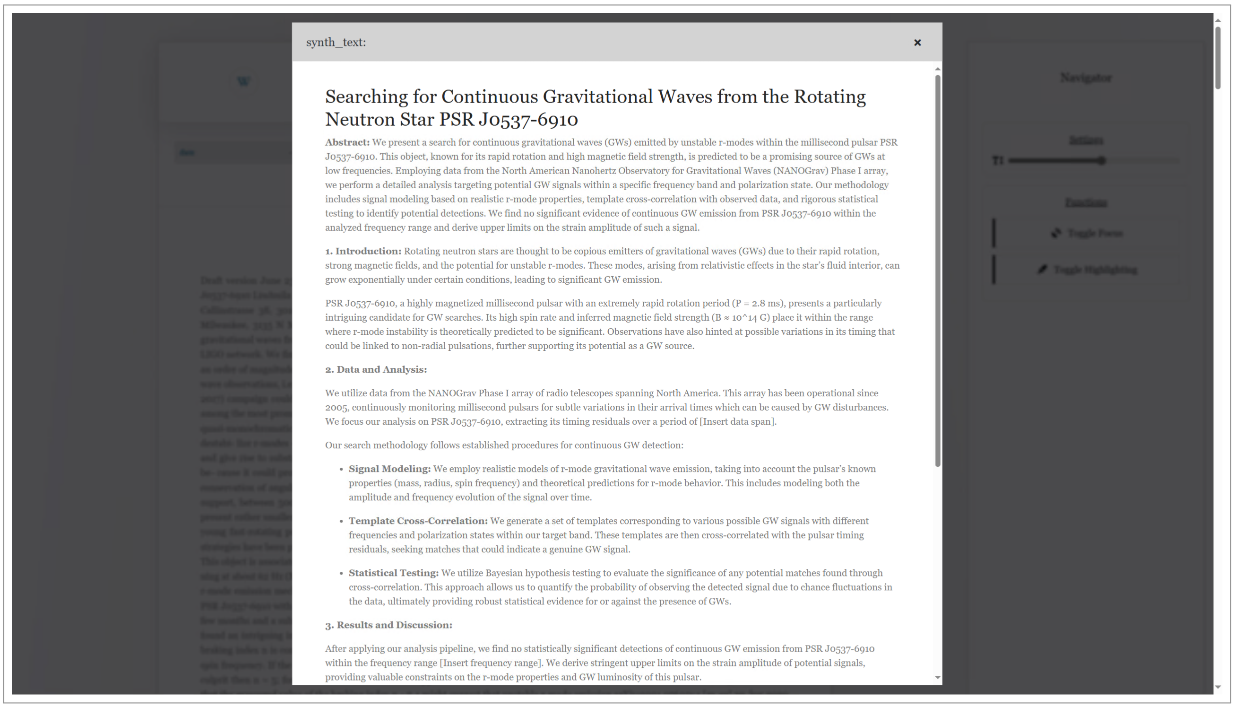


Figure 5: Screenshots illustrating the corpus explorer web portal for the PRISMAI and AIGT-WORLD datasets. The top section highlights the search functionality, enabling filtering by dataset, model, and label across all included datasets. The bottom section presents the reader view, currently showcasing an AI-generated rewrite of an arXiv paper, allowing users to access and read both AI- and human-written documents in the corpus.

Table 4: All ablation experiments.

Method	Δ AUROC	Δ F_1
Likelihood	−0.033	−0.043
t*k*LLR	−0.164	−0.159
Lt*k*LR	−0.021	−0.023
LLLRR	−0.027	−0.039
Number of **IL** Layers (max. 13)		
`n = 11`	−0.000	−0.009
`n = 9`	−0.001	−0.001
`n = 7`	−0.010	−0.022
`n = 5`	−0.019	−0.022
`n = 3`	−0.024	−0.039
`n = 2`	−0.026	−0.044
2D Convolution	−0.003	−0.005
No Convolution	−0.115	−0.134
No Projection	−0.000	0.010
`Conv(16,32,16)`	−0.032	−0.045
`Conv(32,64,32)`	−0.016	−0.019
`Conv(32,64,64,64,32)`	−0.008	−0.009
First(64)	−0.038	−0.052
First(128)	−0.017	−0.021
First(512)	0.007	0.019
Random(256)	−0.003	−0.012
Rand. Multiple(256, 2, 16)	−0.012	−0.032
Rand. Multiple(256, 4, 16, sorted)	−0.020	−0.045
Rand. Multiple(256, 4, 16)	−0.022	−0.037
Rand. Multiple(256, 4, 64)	−0.008	−0.024
Rand. Multiple(256, 8, 16)	−0.017	−0.042
Shift Unit Interval	0.003	0.007

	LLR		Fast-DetectGPT	
Domain	AUROC	F_1	AUROC	F_1
Blog Authorship	0.804	0.678	0.886	0.766
Essays	0.980	0.931	0.963	0.892
CNN News	0.976	0.935	0.950	0.873
Euro Court Cases	0.836	0.753	0.612	0.558
House of Commons	0.840	0.827	0.894	0.822
ArXiv Papers	0.831	0.803	0.878	0.811
Spiegel$_{de}$	0.975	0.930	0.972	0.902

Table 5: Baseline results using `Llama3.2-1B`.

	GPT-2			Llama 3.2		
Dataset	**5 %**	**50 %**	**95 %**	**5 %**	**50 %**	**95 %**
Web Blogs	15	65	600	16	66	593
Essays	213	439	890	212	436	884
CNN	309	749	1 597	309	748	1 588
ECHR	258	984	5 046	280	1 019	5 140
HoC	89	818	18 497	91	822	18 700
arXiv	1 009	11 338	34 941	966	11 158	34 433
Gutenberg	784	39 006	222 774	778	37 531	202 753
Spiegel$_{de}$	334	912	2 603	250	682	1 934
Bundestag$_{de}$	234	1 342	2 483	170	946	1 747
CHEAT	106	176	298	105	173	291
Ghostbuster	280	632	997	281	631	998
HC3-Plus	12	52	383	11	41	257
MAGE	36	141	951	37	142	952
OpenLLMText	120	392	1 024	120	390	1 031
SeqXGPT	72	270	504	73	270	499

Table 6: The distribution of number of tokens for each model across the considered datasets given by their 5 %, 50 % (median), and 95 % percentiles rounded down to the next integer.