Curated Data does not mean Representative Data when training Large Language Models: an Experiment using Representative Data for Italian

Fabio Tamburini[1,*]

[1]FICLIT - University of Bologna, via Zamboni, 32, 40126, Bologna, Italy

Abstract

It is widely accepted in literature that data curation is the first step for a successful pretraining of Large, and Small, Language Models (LLMs). Datasets generally fall into two categories: open datasets are publicly available, fostering transparency, reproducibility, and community-driven improvement, but they often face limitations in scale, diversity, and quality. Closed datasets, typically curated by private entities, can offer greater scale, higher quality, and proprietary data sources, yet they raise concerns around transparency, bias auditing, and public accountability.

This paper presents an experiment aimed at quantitatively measuring the improvements provided by representative datasets for LLM pretraining. We pretrained two small LLMs under the same experimental conditions as the corresponding Italian reference models from the Minerva family, evaluated their performance on standard benchmarks, and used LLM-as-a-Judge to assess the Fluency, Coherence, and Relevance of generated texts on specific tasks. The results support the idea that, while open science and open datasets are important goals, representative corpora, even if closed, are more suitable for LLM pretraining, as they enable better performance under identical experimental conditions.

Keywords

LLM pretraining, representative corpora, text generation evaluation, LLM-as-a-judge

1. Introduction

Large language models (LLMs) have emerged as foundational tools in Natural Language Processing (NLP), powering a wide array of applications from question answering and summarisation to code generation and scientific discovery. Their performance, generalisation ability, and alignment with human values are deeply influenced by the quality, diversity, and scale of the data used during pretraining [1, 2]. As models grow larger and more capable, the need for rigorous data curation practices becomes increasingly critical not only to enhance downstream performance but also to mitigate harmful biases, hallucinations, and environmental costs [3, 4].

Data curation for LLMs involves the collection, filtering, deduplication, classification, and documentation of large-scale textual corpora. These processes aim to balance scale with quality by removing low-signal, harmful, or irrelevant content while preserving linguistic diversity and domain coverage [5, 6]. More recent efforts have highlighted that indiscriminate use of web-scale data may result in the propagation of social biases and misinformation [7], emphasising the importance of carefully designed curation pipelines that consider ethical and societal dimensions [8].

While early work relied heavily on broad, minimally filtered internet scrapes (e.g., Common Crawl), more recent approaches have shifted toward structured, transparent, and task-specific datasets, often constructed through a combination of automated and manual filtering techniques [9]. These developments reflect a growing recognition that model capabilities and behaviours are closely tied to the provenance and properties of their training data. However, the field still lacks standardised methodologies and benchmarks for evaluating curated datasets, presenting challenges for reproducibility and comparative analysis.

1.1. Open vs. Closed Pretraining Datasets

The growing ecosystem of LLMs has revealed a sharp divide between open and closed approaches to data curation. On one hand, open-source initiatives such as BLOOM [10], OPT [2], Pythia [11] and Minerva [12] have committed to full transparency by using publicly available datasets and releasing detailed documentation of their training corpora. These efforts aim to promote reproducibility, community-driven auditing, and equitable access to foundation models. On the other hand, leading commercial models such as GPT-4, Claude and Gemini rely on proprietary or undisclosed datasets, raising questions about accountability, data provenance, and research reproducibility.

The open-data approach is grounded in scientific ideals of transparency and collaborative validation. Models like

CLiC-it 2025: Eleventh Italian Conference on Computational Linguistics, September 24 — 26, 2025, Cagliari, Italy

[*] Corresponding author.

fabio.tamburini@unibo.it (F. Tamburini)
http://corpora.ficlit.unibo.it/People/Tamburini/ (F. Tamburini)
0000-0001-7950-0347 (F. Tamburini)

BLOOM, trained exclusively on open-access sources including multilingual Common Crawl, Project Gutenberg, and academic corpora, exemplify an effort to democratise LLM research and foster global participation [10]. The open release of datasets enables systematic study of data quality, bias, duplication, and domain representation, and it supports downstream development of safer and more equitable AI systems.

In contrast, closed models often cite competitive, ethical, or legal reasons for withholding training data details. OpenAI's GPT-4 report, for example, states that "given the competitive landscape and the safety implications of large-scale models," they have opted not to disclose training data sources. While this protects proprietary advantages and potentially prevents misuse of harmful content, it also hinders external audits of data quality, bias, and copyright compliance. Without transparency, it becomes difficult to evaluate how model performance or behaviour may be influenced by specific sources or omissions.

This divergence has implications for the broader AI research community. The lack of visibility into proprietary datasets exacerbates the reproducibility crisis in machine learning and limits efforts to assess environmental and social impacts of training practices. Conversely, open models, while more transparent, often contend with limitations in data scope and quality due to the exclusion of copyrighted or paywalled content, potentially affecting their competitiveness in knowledge-rich domains.

Ultimately, the tension between open and closed data paradigms reflects competing priorities in the development of foundation models: openness and accountability versus competitive advantage and scalability.

1.2. Key Open Datasets for LLM Pretraining

A number of high-quality, publicly available datasets have become foundational to the training of open-source large language models. These datasets vary in terms of domain coverage, linguistic diversity, and preprocessing strategies, but collectively represent the backbone of transparent and reproducible LLM development.

The Pile [5], a curated 825 GB dataset designed for training language models, combines diverse sources such as academic articles (arXiv), code (GitHub), books, legal documents, and forums to maximise domain coverage. C4 (Colossal Clean Crawled Corpus) [4] is a large-scale, filtered dataset derived from Common Crawl. It removes boilerplate, duplicates, and low-quality text to provide a clean, general-purpose corpus for language modeling. RedPajama [13] presents a reproducible, open alternative to the Llama pretraining dataset. It aggregates content from Common Crawl, Wikipedia, ArXiv, StackExchange, and more, with a focus on transparency and reproducibility. RefinedWeb [14] features a deduplicated and quality-filtered web dataset used to train models such as Falcon. It emphasises a scalable yet high-signal alternative to raw web scrapes. CulturaX [15] is large-scale multilingual web dataset covering 167 languages, designed to improve the cultural and linguistic diversity of LLMs. CulturaX emphasises inclusion of underrepresented languages by sourcing and curating high-quality content from Wikipedia, government websites, and news sources. Books3 (from The Pile) is large collection of digitised books, providing long-form narrative and expository text. Despite its utility, its inclusion has sparked debate due to copyright concerns, underscoring the need for clearer data usage norms.

These datasets are frequently combined or customised depending on the training goals, whether for general-purpose models, multilingual capability, or domain-specific LLMs. CulturaX, in particular, represents a growing movement toward linguistic equity and cultural inclusivity in large-scale model pretraining.

The effort to create open datasets for LLM pretraining that cover a wide range of data inevitably encounters a major challenge: whether or not to include text types that are not freely available on the web. In our view, this is a critical issue when comparing LLMs trained on open data with their counterparts developed by large tech companies using closed datasets, which undoubtedly include a richer and more representative variety of document types for the language or languages being studied. The central concept here is representativeness, which Egbert et al. [16] define as "the extent to which a corpus permits accurate generalisations about the target domain, which involves two components: the extent to which the corpus includes the full range of both text types and linguistic distributions in a domain". In essence, a representative corpus should serve as a statistically valid sample of the population of texts corresponding to the language variety under investigation.

Another point regards the quality of texts published on the Web when compared with curated and edited texts issued by professional publishers. Web texts and published texts differ significantly in form, purpose, authorship, and audience engagement. Web texts, such as blog posts, social media updates, and news articles, tend to be dynamic, hyperlinked, and frequently updated. They emphasise immediacy, brevity, and interactivity, often written in an informal tone to encourage user engagement [17]. In contrast, published texts like academic articles, books, and journals are typically static, peer-reviewed, and follow rigorous editorial standards. These texts prioritise depth, permanence, and formal structure. Additionally, while published texts aim for scholarly credibility and longevity, Web texts often prioritise accessibility, shareability, and multimedia integration. Understanding these distinctions is critical for analysing digital literacy and communication strategies in the information age and, in our opinion, is also critical for pretraining LLM providing "good" and "reliable" texts for teaching a language to a LLM.

This paper aims at exploring and quantify the differences in training a LLM either with open Web data or on a representative corpus examining if the two settings produce some differences in LLM performance, taking contemporary Italian as the reference language.

2. A Representative Dataset

Given the objective of this study, we introduce the reference corpus for contemporary Italian which we use as a template for building the representative corpus employed in our experiments.

2.1. The CORIS Italian Corpus

CORIS design was started in 1998 with the purpose of creating a representative, synchronic, general reference corpus of written contemporary Italian which would be easily accessible and user-friendly [18, 19]. CORIS currently contains 165 million words and has been updated every three years by means of a monitor corpus [20]. It consists of a collection of authentic and commonly occurring texts in electronic format chosen by virtue of their representativeness of contemporary Italian.

After a long design process devoted to a careful definition of relevant textual macro-varieties and their proportions, CORIS has been structured as outlined in Table 1: the largest section, namely 'Press', contains newspapers and periodicals articles, 'Fiction' a collection of novels and short stories while scientific texts and legal/bureaucratic documents where included, respectively, in 'Academic Prose' and 'L&A Prose'. The last two sections contain respectively documents not belonging to the previous categories and texts belonging to Internet language (mainly posts from high quality blogs).

CORIS Section	Proportion
Press	38%
Fiction	25%
Academic Prose	12%
Legal & Admin. Prose	10%
Miscellanea	10%
Ephemera	5%

Table 1
CORIS Sections and their proportions.

Based on the general CORIS schema outlined in Table 1, we created an 11.6 billion-token corpus that includes the same textual macro-varieties and maintains the same CORIS balancing. This corpus was constructed by selecting materials from the previously mentioned CulturaX project and incorporating large sections of specific published texts. This extensive, curated, and representative training corpus was then used to train our new "CORIS-llm" language model, following the procedure described in the next section.

3. Experiment Settings

3.1. LLM pretraining

Minerva is the first family of LLMs pretrained from scratch on Italian [12] and emerged as a standard reference for Italian NLP. A prior study pretrained an Italian model based on GPT-2 from scratch [21], but it used a relatively small 117M-parameter set, making it not directly comparable to modern LLMs or the more recent Minerva family.

In order to perform a fair comparison with the Minerva models, we adopted exactly the same pretraining settings and hyperparameters described in [12]. We pretrained the models using the MosaicML LLM-Foundry[1] package concentrating our efforts on two models: a 350M-parameter model trained on a single node equipped with four A100-64GB GPUs for an equivalent number of steps as the Minerva-350M model and 1B parameter model trained on 2 nodes in the same way as Minerva-1B[2]. While a 11.6 billion-token corpus is big enough for pretraining a 350M model, it is too small, following the Chinchilla rule [22] involving a parameter/token ratio of 1:20, for a 1B model, thus, in this second case, we could expect some performance degradation.

A detailed quality analysis of the Minerva dataset is contained in the original paper [12].

3.2. First Evaluation on Standard Benchmarks

The evaluation of LLMs has traditionally relied on a suite of standardised benchmarks designed to assess a broad range of linguistic, reasoning, and task-specific capabilities. These benchmarks enable systematic comparison across models and facilitate progress tracking in natural language processing.

To address the need for evaluating generation-based tasks, LAMBADA [23] tests a model's ability to predict the final word of a passage based on broad context, emphasising long-range dependency modelling. In parallel, benchmarks such as WinoGrande [24] and HellaSwag [25] target common-sense reasoning and disambigua-

[1] https://github.com/mosaicml/llm-foundry
[2] https://huggingface.co/sapienzanlp/Minerva-XXX-base-v1.0

Model	ARC-C	ARC-E	BoolQ	GSM8K	HS	MMLU	PIQA	SciQ	TQA	WG	AVG
Minerva-350M-base-v1.0 [12]	24.6	36.4	60.7	48.2	32.6	25.7	59.5	63.7	46.5	58.4	45.6
Minerva-350M-base-v1.0 (our)	24.7	36.4	60.7	48.4	32.6	25.7	59.0	55.0	46.6	56.1	44.5
CORISllm-350M-base	25.1	34.9	49.3	47.0	31.9	25.6	57.4	52.3	46.7	57.1	42.7
Minerva-1B-base-v1.0 [12]	26.6	42.2	57.1	49.7	39.6	27.0	62.9	73.5	44.6	60.0	48.3
Minerva-1B-base-v1.0 (our)	26.9	42.2	54.3	49.5	39.5	27.1	63.0	65.6	44.6	59.1	47.2
CORISllm-1B-base	26.1	36.5	44.7	47.4	35.8	26.9	60.7	53.9	46.6	56.8	43.5

Table 2

ITA-Bench Evaluation results for CORISllm-350M, CORISllm-1B and the corresponding Minerva models. "(our)" indicates the recalculation of Minerva performance made by us fixing the random seed for a fair comparison with CORISllm models.

tion, probing a model's depth of understanding beyond surface-level patterns.

More recently, MMLU (Massive Multitask Language Understanding) has been introduced as a collaborative effort to assess a wide range of LLM competencies ranging from law and medicine to physics and philosophy offering a broad-spectrum evaluation across 57 subjects to test a model's ability to generalise across domains [26].

While existing evaluation benchmarks are highly valuable, they are primarily designed to assess LLM performance in English and are therefore not suitable for our purposes. Recently, a group of Italian researchers introduced a promising new benchmark, called ITA-Bench, for evaluating LLMs in Italian. This suite combines automatically translated versions of popular English benchmarks with adapted, manually curated datasets for Italian [27]. We adopted ITA-Bench for the initial evaluation of our new LLMs and conducted a preliminary comparison with an equivalent Minerva model.

Table 2 presents the results of CORISllm-350M and CORISllm-1B on ITA-Bench, alongside a comparison with the corresponding Minerva models. Overall, the two LLMs demonstrate comparable performance: Minerva performs better on certain tasks, while CORISllm slightly outperforms it on others.

On average, the Minerva models show slightly better performance; however, these results must be interpreted in light of the nature of the benchmark. ITA-Bench focuses primarily on tasks involving commonsense reasoning and scientific knowledge retrieval, which are not well-suited for assessing differences in text generation capabilities. Pretraining an LLM on a representative corpus does not inherently confer an advantage in reasoning or STEM-related tasks because the dataset used for pretraining does not contain specific materials useful for increasing performance on STEM-related tasks and no specific methods were used to promote the development of reasoning abilities. Accordingly, CORISllm and Minerva perform similarly on ITA-Bench. To properly evaluate our research hypothesis, a more targeted assessment of text generation abilities is required.

3.3. Text Generation Quality Evaluation

Evaluating LLM-generated texts is inherently challenging, and assessing the quality of these textual outputs is even more complex [28].

Our primary objective is to conduct a careful evaluation of the quality of texts generated by LLMs. Specifically, we aim to compare an LLM trained on "open" but non-representative datasets, namely the Minerva family, with one trained on a representative and balanced dataset, CORISllm. The comparison focuses on commonly used human evaluation metrics: *Fluency*, internal *Coherence* and text *Relevance* to the given task.

To ensure a fair evaluation, it is necessary to generate and assess a substantial number of texts. For this purpose, we adopted the LLM-as-a-Judge (LaaJ) approach, after a comparison of LLMs annotations with human judgments.

We designed six distinct prompts, each corresponding to one of the six CORIS macrovarieties: a short newspaper article, a children's fairy tale, an abstract of a scientific paper, a judgment for a crime, a trip description, and a brief movie review, and generated 50 outputs each. Table 3 presents the prompts used to stimulate the LLMs to generate texts.

The following sections first describe the human evaluation process, followed by the LaaJ methodology we employed to achieve our objective.

3.3.1. Human Evaluation of LLM Outputs

Human evaluation remains the gold standard for assessing the quality of natural language outputs produced by LLMs. Despite the growing sophistication of automated metrics and model-based evaluators, human judgments are uniquely capable of capturing nuanced dimensions of quality such as contextual appropriateness, subtle coherence errors, pragmatic relevance, and factual accuracy. Consequently, human assessments are widely used in both benchmarking LLMs and validating automatic evaluation methods.

Human evaluation of LLM outputs is typically carried out using either rating scales (e.g., Likert scales), pairwise comparisons, or ranking protocols. Each approach has strengths and limitations: scalar ratings allow fine-grained feedback but may suffer from rater calibration issues, while relative comparisons often yield more consistent judgments.

In the context of LLM outputs, common evaluation criteria include fluency, coherence, relevance, factual accuracy, and harmlessness or bias. For instance, the HELM benchmark [29] employs extensive human annotation pipelines to assess these aspects. Fluency is often reliably judged, but tasks like evaluating factual consistency or detecting hallucinations present greater challenges. Human annotators are also crucial for detecting subtle harms, such as stereotyping or toxicity, which automated tools frequently miss or misclassify [3].

Despite its value, human evaluation has notable limitations. It is expensive, time-consuming, and subject to inter-rater variability, which can obscure subtle differences between systems. Additionally, annotator background and task framing can influence outcomes. For example, work has shown that crowdworker evaluations can differ systematically from domain-expert judgments, particularly on complex tasks like summarisation or question answering [30].

To compare the behaviour of the considered LaaJ systems with human judgments, we conducted a small experiment where three expert linguists manually evaluated 120 texts produced by Minerva-350M in response to the six prompts given in Table 3. The annotators were asked to evaluate the LLM-generated texts according to the three selected metrics: the instructions given to them was almost identical to the prompts in Tables 8, 9 and 10 we used for LaaJ. Table 4 (top-left section) shows the Spearman Rank Correlation Coefficients (SRCC) between the rankings provided by the three human annotators A1-A3, who assigned scores on a 5-point Likert scale. The correlations were relatively low, highlighting the challenges human annotators face in consistently grading text production using similar criteria.

Due to the low correlations observed, particularly in the assessment of Fluency, we decided against using the human annotations to calibrate our LaaJ systems and chose to rely solely on the LaaJ methodology for the evaluations.

Text Macro-variety	**Prompt**
Press	*"Scrivi un articolo di quotidiano su un fatto di cronaca inventato composto al massimo da cinque frasi.\n\n Ieri pomeriggio"*
Fiction	*"Inventa una piccola favola per bambini composta al massimo da cinque frasi.\n\n C'era una volta"*
Academic Prose	*"Scrivi un sommario, o abstract, di un articolo scientifico composto al massimo da cinque frasi.\n\n In questo articolo"*
Legal & Admin. Prose	*"Scrivi una sentenza di condanna per un piccolo furto composta al massimo da cinque frasi.\n\n Questa sezione penale"*
Miscellanea	*"Descrivi un viaggio in un posto qualsiasi utilizzando al massimo cinque frasi.\n\n Lo scorso anno ho visitato"*
Ephemera	*"Scrivi una recensione su un film composta al massimo da cinque frasi.\n\n Il film"*

Table 3
Prompts used for generating texts by CORISllm and Minerva.

	A2	**A3**	**Llama**	**Gem2**
A1				
Flu.	.312	.152	.316	.315
Coh.	.470	.542	.566	.446
Rel.	.465	.761	.558	.586
A2				
Flu.	-	.428	.210	.069
Coh.	-	.551	.324	.317
Rel.	-	.476	.412	.389
A3				
Flu.	-	-	.156	.042
Coh.	-	-	.350	.262
Rel.	-	-	.638	.564
Llama				
Flu.	-	-	-	.630
Coh.	-	-	-	.597
Rel.	-	-	-	.613

Table 4
Spearman Rank Correlation Coefficients between the three human annotators (A1-A3) and the two LaaJ (Llama-3.3-70B and Gemini-2.0-flash).

3.3.2. LLMs as Automated Judges of Text Quality

Recent advances LLMs have opened new avenues for evaluating textual outputs in NLP. Traditionally, the evaluation of text generation has relied heavily on human judgments, which, while high in fidelity, are costly, time-consuming, and often inconsistent due to inter-annotator variability [31]. In contrast, LLMs such as GPT-3/4, Palm and Gemini have demonstrated potential not only in generating text but also in providing reliable meta-judgments about language quality, including fluency, coherence, and relevance.

Several studies have investigated the reliability of LLMs as automatic evaluators. For instance, G-Eval [32] highlights that LLMs can approximate human judgments in multi-dimensional evaluation tasks when properly prompted. As shown in the nice review by Li et al. [33], it is possible to set up a framework where an LLM acts as a zero-shot or few-shot judge, providing ordinal or scalar ratings that correlate highly with human annotations. This correlation is particularly strong when the models are instructed explicitly to focus on specific dimensions of quality, such as grammatical fluency or semantic relevance.

In terms of **Fluency**, LLMs have internalised extensive grammatical structures through pretraining on large corpora, enabling them to effectively recognise and assess grammaticality and naturalness. For **Coherence**, models evaluate the logical consistency and flow of ideas across sentences or turns, especially when equipped with context windows that span multiple paragraphs. Evaluating **Relevance**, the alignment of a response to a prompt or topic, has also been shown to benefit from LLMs' contextual awareness and knowledge grounding.

In summary, LLMs have emerged as credible tools for evaluating textual quality across multiple dimensions: when applied with careful prompt design and interpretative caution, they can serve as scalable, cost-effective complements to human assessment.

In order to avoid any inconsistency introduced by human judgments, we decided to rely only on two different LLMs for evaluating the quality of texts produced by CORISllm and Minerva models.

We adopted a powerful online LLM, namely *Gemini-2.0-flash* through Google APIs, and an offline, quantised model, namely *bartowski/Llama-3.3-70B-Instruct-Q6_K_L* downloaded from the Huggingface repository[3].

Tables 8, 9 and 10 show the three prompts we have designed for asking the two LaaJ to evaluate, using a 5-point Likert scale, *Fluency*, *Coherence* and *Relevance* of the texts generated by CORISllm-350M/1B and Minerva-350M/1B. For designing these prompt we took inspiration from similar prompts proposed in G-Eval [32]. The separators '##SYSTEM##', '##USER##' and '##ASSISTANT##' for marking the three different blocks of information in the prompts were replaced with empty lines for Gemini prompts and with the appropriate separators for prompts proposed to the Llama judge.

To assess the reliability of their judgments, we first evaluated the agreement between the two LaaJ systems and the human annotators. Table 4 also reports the SRCC between each LaaJ and the human annotators. While the two LaaJ systems show high mutual correlation, their agreement with individual human annotators is lower, though still comparable to the level of agreement observed between human annotators themselves. This further supports the case for favoring LaaJ-generated annotations over those produced by humans.

Table 5 shows the (SRCCs) between Gemini-2.0-flash and the quantised Llama-3.3-70B judges when evaluating 600 new texts produced by CORISllm and Minerva models (300 for each model). Correlations are all quite high and highly significant, thus we can reliably use these automatic judges for evaluating the textual production of the tested models.

Model	SRCC	p-value
CORISllm-350M		
Flu	0.7178	⋘0.001
Coh	0.6369	⋘0.001
Rel	0.7036	⋘0.001
CORISllm-1B		
Flu	0.6462	⋘0.001
Coh	0.6957	⋘0.001
Rel	0.7169	⋘0.001
Minerva-350M		
Flu	0.6576	⋘0.001
Coh	0.6654	⋘0.001
Rel	0.7048	⋘0.001
Minerva-1B		
Flu	0.5844	⋘0.001
Coh	0.6640	⋘0.001
Rel	0.7235	⋘0.001

Table 5
SRCCs between the LLM judges Gemini-2.0-flash and quantised LLama-3.3-70B when evaluating the 600 texts produced by CORISllm and Minerva (300 for each model).

4. Results

Tables 6 and 7 present the means and standard deviations of the scores assigned by the two judges across the three evaluation metrics to the 600 texts forming the evaluation dataset. The tables also display the results of a t-test for independent samples, which assesses the statistical significance of the differences in means.

Examining the evaluations provided by Gemini, we observe a notable increase in the scores assigned to CORISllm-350M compared to the equivalent Minerva-350M model. Furthermore, the scores for it are so high that they are comparable, and not significantly different, to those of the larger Minerva-1B model, with the exception of the Relevance metric. With regard to CORISllm-1B, it performs much better than Minerva-350M, as expected, and more or less on par with Minerva-1B, exhibiting better performance on Fluency and worse on Relevance. All differences that are statistically significant are indicated by the asterisks next to the metric.

Regarding the Llama-3.3-70B judge, CORISllm-350M consistently receives significantly higher scores than the equivalent Minerva-350M model, and its scores are comparable to those of the Minerva-1B model across all metrics. Using this judge, CORISllm-1B performs much better than both Minerva models in a highly significant way,

[3] https://huggingface.co/bartowski/Llama-3.3-70B-Instruct-GGUF

	Minerva-350M	**Minerva-1B**
	Flu=2.49±1.02	Flu=2.83±1.08
	Coh=1.77±0.80	Coh=2.1±0.97
	Rel=1.73±1.26	Rel=2.34±1.59
CORISllm-350M	←	↑
Flu=2.74±0.96	Flu**	~~Flu~~
Coh=2.01±0.80	Coh***	~~Coh~~
Rel=1.97±1.39	Rel*	Rel**
CORISllm-1B	←	
Flu=3.1±0.98	Flu***	Flu** ←
Coh=2.18±0.92	Coh***	~~Coh~~
Rel=1.95±1.35	Rel*	Rel*** ↑

Table 6
Means and standard deviations of the scores assigned by Gemini-2.0-flash to the texts produced by the tested models. Stars near the metric abbreviations indicate the t-test significance (~~X~~-not sig., *-sig., **-very sig., ***-highly sig.). Arrows indicate which model performs best.

	Minerva-350M	**Minerva-1B**
	Flu=2.08±1.04	Flu=2.28±1.17
	Coh=1.50±0.91	Coh=1.76±1.49
	Rel=1.60±1.05	Rel=2.07±1.39
CORISllm-350M	←	-
Flu=2.49±1.01	Flu***	~~Flu~~
Coh=1.73±1.01	Coh**	~~Coh~~
Rel=1.98±1.20	Rel***	~~Rel~~
CORISllm-1B	←	←
Flu=2.80±1.03	Flu***	Flu***
Coh=2.04±1.19	Coh***	Coh**
Rel=2.10±1.31	Rel***	~~Rel~~

Table 7
Means and standard deviations of the scores assigned by the quantised Llama-3.3-70B to the texts produced by the tested models. Stars near the metric abbreviations indicate the t-test significance (~~X~~-not sig., *-sig., **-very sig., ***-highly sig.). Arrows indicate which model performs best.

except for the Relevance metric when compared with Minerva-1B for which there seems to be no significant differences.

##SYSTEM##
Tu sei un linguista esperto nella valutazione dei testi. Ti verrà fornita la descrizione di un esercizio e lo svolgimento di questo esercizio da parte di un'AI.
Il tuo compito è valutare lo svolgimento in base a una metrica. Assicurati di leggere e comprendere attentamente queste istruzioni. Tieni aperto questo documento durante la revisione e consultalo quando necessario.
Criteri di valutazione:
Coerenza (1-5): la qualità globale di tutte le frasi. Il testo dovrebbe essere ben strutturato e ben organizzato. Il testo non dovrebbe contenere solo un mucchio di informazioni correlate, ma dovrebbe svilupparsi da una frase a un corpo coerente di informazioni su un argomento.
Fluenza (1-5): la qualità dello svolgimento in termini di grammatica, ortografia, punteggiatura, scelta delle parole e struttura delle frasi.
- 1/2. Scarsa. Lo svolgimento presenta molti errori che lo rendono difficile da comprendere o lo rendono poco naturale.
- 3. Lo svolgimento presenta alcuni errori che compromettono la chiarezza o la scorrevolezza del testo, ma i punti principali sono comunque comprensibili. - 4. Buona. Lo svolgimento presenta pochi errori ed è facile da leggere e seguire. - 5. Ottima. Lo svolgimento non contiene errori ed è facile da leggere e seguire.
Fasi di valutazione:
1. Leggi attentamente lo svolgimento e identifica gli errori grammaticali, ortografici e sintattici. 2. Assegna un punteggio per la fluenza su una scala da 1 a 5, dove 1 è il punteggio più basso e 5 il punteggio più alto in base ai Criteri di valutazione.

##USER##
Descrizione dell'esercizio:
{{Esercizio}}
Svolgimento:
{{Svolgimento}}

##ASSISTANT##
Modulo di valutazione (SOLO punteggi):
Fluenza:

Table 8
Prompts for LaaJ ranking Fluency. '{{Esercizio}}' and '{{Svolgimento}}' were replaced respectively by the description of the task and the text produced by the tested LLM.

5. Discussion & Conclusions

In this study, we examined how the choice of data for LLM pretraining affects performance, emphasizing the importance of using a representative corpus to enhance the quality of text produced by generative LLMs.

Using the design framework of the CORIS corpus, a representative corpus of contemporary Italian, we pretrained two LLMs following exactly the same process used for the Minerva models [12]. However, instead of the original dataset, we used a new 11.6 billion-token representative corpus specifically structured to align with the CORIS macrovarieties.

When evaluating the textual production of equivalent models across Fluency, internal Coherence, and Relevance to the assigned task, CORISllm outperformed Minerva. Due to the limited dimensions of the training corpus, suitable to pretrain 350M models and less 1B models, the results are more neat on smaller models. In any case, this points in the direction that using representative and balanced corpora for LLM pretraining has an impact on performance. In our experiments, CORISllm-350M, despite having only one-third of the model parameters, performed nearly on par with Minerva-1B in terms of generative text quality.

##SYSTEM##
Tu sei un linguista esperto nella valutazione dei testi. Ti verrà fornita la descrizione di un esercizio e lo svolgimento di questo esercizio da parte di un’AI.
Il tuo compito è valutare lo svolgimento in base a una metrica. Assicurati di leggere e comprendere attentamente queste istruzioni. Tieni aperto questo documento durante la revisione e consultalo quando necessario.
Criteri di valutazione:
Coerenza (1-5): la qualità globale di tutte le frasi. Il testo dovrebbe essere ben strutturato e ben organizzato. Il testo non dovrebbe contenere solo un mucchio di informazioni correlate, ma dovrebbe svilupparsi da una frase a un corpo coerente di informazioni su un argomento.
Fasi di valutazione:
1. Leggi attentamente lo svolgimento e identifica l’argomento principale e i punti chiave. 2. Analizza il contenuto di ogni frase e valuta se frasi successive sono legate logicamente e strutturalmente. 3. Assegna un punteggio per la coerenza su una scala da 1 a 5, dove 1 è il punteggio più basso e 5 il punteggio più alto in base ai Criteri di valutazione.

##USER##
Descrizione dell’esercizio:
{{Esercizio}}
Svolgimento:
{{Svolgimento}}

##ASSISTANT##
Modulo di valutazione (SOLO punteggi):
Coerenza:

Table 9
Prompts for LaaJ ranking Coherence. ‘{{Esercizio}}’ and ‘{{Svolgimento}}’ were replaced respectively by the description of the task and the text produced by the tested LLM.

The goal of this work was not to create a complete family of LLMs pretrained on representative corpora and ready for production deployment. Rather, we aimed to provide a proof-of-concept study that emphasises the need for greater attention to training corpora in order to develop better models.

While the openness of training data is certainly a valuable principle, the results presented here suggest that it is equally important to incorporate high-quality published texts into the training process in order to enhance performance without altering the transformer model. Since such materials are often protected by copyright, it is essential to establish specific agreements with publishers.

Due to copyright restrictions on portions of our pretraining corpus, we are unable to distribute it freely. CORISllm models are available upon request.

##SYSTEM##
Tu sei un linguista esperto nella valutazione dei testi. Ti verrà fornita la descrizione di un esercizio e lo svolgimento di questo esercizio da parte di un’AI.
Il tuo compito è valutare lo svolgimento in base a una metrica. Assicurati di leggere e comprendere attentamente queste istruzioni. Tieni aperto questo documento durante la revisione e consultalo quando necessario.
Criteri di valutazione:
Rilevanza (1-5): Lo svolgimento deve includere solo informazioni allineate con la descrizione dell’esercizio. Dovrai penalizzare gli svolgimenti che contengono informazioni o argomenti non rilevanti rispetto alla descrizione.
Fasi di valutazione:
1. Leggi attentamente lo svolgimento e identifica l’argomento principale e i punti chiave. 2. Confronta lo svolgimento con la descrizione dell’esercizio. 3. Assegna un punteggio di rilevanza da 1 a 5, dove 1 è il punteggio più basso e 5 il punteggio più alto in base ai Criteri di valutazione.

##USER##
Descrizione dell’esercizio:
{{Esercizio}}
Svolgimento:
{{Svolgimento}}

##ASSISTANT##
Modulo di valutazione (SOLO punteggi):
Rilevanza:

Table 10
Prompts for LaaJ ranking Relevance. ‘{{Esercizio}}’ and ‘{{Svolgimento}}’ were replaced respectively by the description of the task and the text produced by the tested LLM.

Acknowledgments

We acknowledge the CINECA[4] award no. HP10CPJEG0 (project DARE4LLM) under the ISCRA initiative, for the availability of HPC resources and support.

References

[1] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, in: Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS ’20, 2020.

[2] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin, T. Mi-

[4] https://www.cineca.it/en

haylov, M. Ott, S. Shleifer, K. Shuster, D. Simig, P. S. Koura, A. Sridhar, T. Wang, L. Zettlemoyer, OPT: Open Pre-trained Transformer Language Models, 2022. arXiv:2205.01068.

[3] E. M. Bender, T. Gebru, A. McMillan-Major, S. Shmitchell, On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?, FAccT '21, Association for Computing Machinery, New York, NY, USA, 2021, p. 610–623.

[4] J. Dodge, M. Sap, A. Marasović, W. Agnew, G. Ilharco, D. Groeneveld, M. Mitchell, M. Gardner, Documenting large webtext corpora: A case study on the colossal clean crawled corpus, in: M.-F. Moens, X. Huang, L. Specia, S. W.-t. Yih (Eds.), Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 1286–1305.

[5] L. Gao, S. Biderman, S. Black, L. Golding, T. Hoppe, C. Foster, J. Phang, H. He, A. Thite, N. Nabeshima, S. Presser, C. Leahy, The Pile: An 800GB Dataset of Diverse Text for Language Modeling, 2020. arXiv:2101.00027.

[6] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, J. Mach. Learn. Res. 21 (2020).

[7] A. Abid, M. Farooqi, J. Zou, Persistent Anti-Muslim Bias in Large Language Models, in: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, Association for Computing Machinery, New York, NY, USA, 2021, p. 298–306.

[8] L. Weidinger, J. Uesato, M. Rauh, C. Griffin, P.-S. Huang, J. Mellor, A. Glaese, M. Cheng, B. Balle, A. Kasirzadeh, C. Biles, S. Brown, Z. Kenton, W. Hawkins, T. Stepleton, A. Birhane, L. A. Hendricks, L. Rimell, W. Isaac, J. Haas, S. Legassick, G. Irving, I. Gabriel, Taxonomy of Risks posed by Language Models, in: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, FAccT '22, Association for Computing Machinery, New York, NY, USA, 2022, p. 214–229.

[9] N. Brandizzi, H. Abdelwahab, A. Bhowmick, L. Helmer, B. J. Stein, P. Denisov, Q. Saleem, M. Fromm, M. Ali, R. Rutmann, F. Naderi, M. S. Agy, A. Schwirjow, F. Küch, L. Hahn, M. Ostendorff, P. O. Suarez, G. Rehm, D. Wegener, N. Flores-Herr, J. Köhler, J. Leveling, Data Processing for the OpenGPT-X Model Family, 2024. arXiv:2410.08800.

[10] T. L. Scao, A. Fan, C. Akiki, E. Pavlick, S. Ilić, D. Hesslow, many others., BLOOM: A 176B-Parameter Open-Access Multilingual Language Model, 2023. arXiv:2211.05100.

[11] S. Biderman, H. Schoelkopf, Q. G. Anthony, H. Bradley, K. O'Brien, E. Hallahan, M. A. Khan, S. Purohit, U. S. Prashanth, E. Raff, A. Skowron, L. Sutawika, O. Van Der Wal, Pythia: A Suite for Analyzing Large Language Models Across Training and Scaling, in: A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, J. Scarlett (Eds.), Proceedings of the 40th International Conference on Machine Learning, volume 202 of *Proceedings of Machine Learning Research*, PMLR, 2023, pp. 2397–2430.

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

[13] M. Weber, D. Y. Fu, Q. Anthony, Y. Oren, S. Adams, A. Alexandrov, X. Lyu, H. Nguyen, X. Yao, V. Adams, B. Athiwaratkun, R. Chalamala, K. Chen, M. Ryabinin, T. Dao, P. Liang, C. Ré, I. Rish, C. Zhang, RedPajama: an Open Dataset for Training Large Language Models, NeurIPS Datasets and Benchmarks Track (2024).

[14] G. Penedo, Q. Malartic, D. Hesslow, R. Cojocaru, H. Alobeidli, A. Cappelli, B. Pannier, E. Almazrouei, J. Launay, The RefinedWeb Dataset for Falcon LLM: Outperforming Curated Corpora with Web Data Only, in: A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, S. Levine (Eds.), Advances in Neural Information Processing Systems, volume 36, Curran Associates, Inc., 2023, pp. 79155–79172.

[15] T. Nguyen, C. V. Nguyen, V. D. Lai, H. Man, N. T. Ngo, F. Dernoncourt, R. A. Rossi, T. H. Nguyen, CulturaX: A cleaned, enormous, and multilingual dataset for large language models in 167 languages, in: N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, N. Xue (Eds.), Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), ELRA and ICCL, Torino, Italia, 2024, pp. 4226–4237.

[16] J. Egbert, D. Biber, B. Gray, Corpus Representativeness: A Conceptual and Methodological Framework, Cambridge University Press, 2022, p. 52–67.

[17] S. C. Herring, Computer-Mediated Discourse Analysis: An Approach to Researching Online Behavior, Learning in Doing: Social, Cognitive and Computational Perspectives, Cambridge University Press, 2004, p. 338–376.

[18] F. Tamburini, I corpora del FICLIT, Università di Bologna: CORIS/CODIS, BoLC e DiaCORIS, in: Proceedings of the LIV Congresso Internazionale di Studi della Società di Linguistica Italiana, 2022,

pp. 189–197.

[19] R. Rossini Favretti, F. Tamburini, C. De Santis, CORIS/CODIS: A corpus of written Italian based on a defined and a dynamic model, in: A. Wilson, P. Rayson, T. McEnery (Eds.), A Rainbow of Corpora: Corpus Linguistics and the Languages of the World, Lincom-Europa, Munich, 2002, pp. 27–38.

[20] J. Sinclair, Corpus, Concordance, Collocation, Oxford University Press, 1991.

[21] L. D. Mattei, M. Cafagna, F. Dell'Orletta, M. Nissim, M. Guerini, Geppetto carves italian into a language model, in: Proceedings of the 7th Italian Conference on Computational Linguistics (CLiC-it 2020), CEUR Workshop Proceedings, Bologna, Italy, 2020.

[22] J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. de Las Casas, L. A. Hendricks, J. Welbl, A. Clark, T. Hennigan, E. Noland, K. Millican, G. van den Driessche, B. Damoc, A. Guy, S. Osindero, K. Simonyan, E. Elsen, O. Vinyals, J. W. Rae, L. Sifre, Training compute-optimal large language models, in: Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS '22, Curran Associates Inc., Red Hook, NY, USA, 2022.

[23] D. Paperno, G. Kruszewski, A. Lazaridou, N. Q. Pham, R. Bernardi, S. Pezzelle, M. Baroni, G. Boleda, R. Fernández, The LAMBADA dataset: Word prediction requiring a broad discourse context, in: K. Erk, N. A. Smith (Eds.), Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Berlin, Germany, 2016, pp. 1525–1534.

[24] K. Sakaguchi, R. Le Bras, C. Bhagavatula, Y. Choi, Winogrande: An adversarial winograd schema challenge at scale, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2020, pp. 8732–8740.

[25] R. Zellers, A. Holtzman, Y. Bisk, A. Farhadi, Y. Choi, HellaSwag: Can a machine really finish your sentence?, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 4791–4800.

[26] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, J. Steinhardt, Measuring massive multitask language understanding, in: Proceedings of the International Conference on Learning Representations (ICLR), 2021.

[27] L. Moroni, S. Conia, F. Martelli, R. Navigli, Itabench: Towards a more comprehensive evaluation for Italian LLMs, in: F. Dell'Orletta, A. Lenci, S. Montemagni, R. Sprugnoli (Eds.), Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024), CEUR Workshop Proceedings, Pisa, Italy, 2024, pp. 584–599.

[28] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang, W. Ye, Y. Zhang, Y. Chang, P. S. Yu, Q. Yang, X. Xie, A survey on evaluation of large language models, ACM Trans. Intell. Syst. Technol. 15 (2024).

[29] P. Liang, R. Bommasani, T. Lee, D. Tsipras, D. Soylu, et al., Holistic evaluation of language models, Transactions on Machine Learning Research (2023).

[30] M. Karpinska, N. Akoury, M. Iyyer, The perils of using Mechanical Turk to evaluate open-ended text generation, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 1265–1285.

[31] C. van der Lee, A. Gatt, E. van Miltenburg, E. Krahmer, Human evaluation of automatically generated text: Current trends and best practice guidelines, Computer Speech & Language 67 (2021) 101151.

[32] Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, C. Zhu, G-Eval: NLG Evaluation using Gpt-4 with Better Human Alignment, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, 2023, pp. 2511–2522.

[33] H. Li, Q. Dong, J. Chen, H. Su, Y. Zhou, Q. Ai, Z. Ye, Y. Liu, LLMs-as-Judges: A Comprehensive Survey on LLM-based Evaluation Methods, 2024. arXiv:2412.05579.