How far does the sequence of compositions impact Multilingual Pre-Training?

Leonardo Ranaldi[1], Giulia Pucci[2] and Fabio Massimo Zanzotto[3]

[1]School of Informatics, University of Edinburgh, UK.
[2]Department of Computing Science, University of Aberdeen, UK.
[3]Università degli Studi Roma "Tor Vergata", Roma, Italy.

Abstract

An Efficient strategy for conducting pre-training of language models is the concatenation of contiguous sequences of text of fixed length through *causal masking* that estimates the probability of each token given its context. Yet earlier work suggests that this technique affects the performance of the model as it might include misleading information from previous text sequences during pre-training. To fill this gap, intra-context and rank-based causal masking techniques have been proposed, in which the probability of each token is conditional only on the previous ones in the same document or ranked sequences, avoiding misleading information from different contexts. However, the sequences provided by the use of these techniques have been little explored, overlooking the opportunity to optimise the composition by manipulating the volume and heterogeneity in the sequences and improving unbalance pre-training settings. In this paper, we demonstrate that organising text chunks based on a policy that aligns with text similarity effectively improve pre-training, enhances the learning and cross-lingual generalisation capabilities of language models, maintains efficiency, and allows for fewer instances.

Keywords

Large Language Models, Pre-training Methods, Cross-lingual Generalisation,

1. Introduction

Large language models (LLMs) are pre-trained on huge amounts of documents by optimizing a language modelling objective and show an intriguing ability to solve various downstream NLP tasks. Ranaldi et al. [1] in multilingual settings and later Zhao et al. [2] highlighted the importance of pre-training data quality, diversity and composition methodologies. Our research takes a step further by exploring the influence of the pre-training sequences heterogeneity for cross-lingual generalisation. This potentially leads to significant advancements in understanding LLMs' learning properties.

In decoder-only architectures pre-training, the constructions of the instances are based on *packing* that combines randomly sampled texts (i.e., documents) into a *chunk* that matches the size of the context window without using any selection policy. Then, the causal masking predicts the next token conditioned on the previous, including those from different documents (portions of non-contiguous texts) in the chunk. The ways to mitigate this arbitrary procedure are: (i) intra-document causal masking [3], where the likelihood of each token is conditioned on the previous from the same document [3] and retrieval-based masking [2] where similar documents retrieved by retrieval systems condition likelihood.

To study the role of heterogeneity and volume of samples in sequence composition strategies (i.e., packing and masking pipelines), we pre-train language models using different masking approaches (described in §2.2) and compare them with models pre-trained via the traditional causal masking with different packing approaches by varying amount of the sequence composition of the documents in the pre-training chunks. Whilst for studying the impact on cross-lingual generalisation we use cross-lingual settings (i.e., Italian English). Complementing the foundation approaches proposed in [1, 2],we operate via bilingual corpora. Hence, we analyse the results produced by a commonly used baseline method that randomly samples and packs documents (`RandomChunk`), a process that samples and packs documents from the same source based on their composition and origin (`UniChunk`), and then operate via efficient retrieval-based packing method, which retrieves and packs related documents (§2.1).

The experimental results indicate that operating via causal masking (`RandomChunk`) with arbitrary sequence patterns of documents leads to the inclusion of misleading information that stems from different context during pre-training (§3), impacting in a negatively the performance of the models in downstream tasks (§4). Instead, intra-document causal masking, which avoids the misleading phenomena during pre-training, significantly improves the models' performance and does not impact the runtime. Although intra-document causal masking performs well, it limits the operability of sequence composition mixing documents from different corpora (in our

CLiC-it 2024: Tenth Italian Conference on Computational Linguistics, Dec 04 — 06, 2024, Pisa, Italy
lranaldi@ed.ac.uk (L. Ranaldi); g.pucci.24@abdn.uk (G. Pucci); fabio.massimo.zanzotto@uniroma2.it (F. M. Zanzotto)

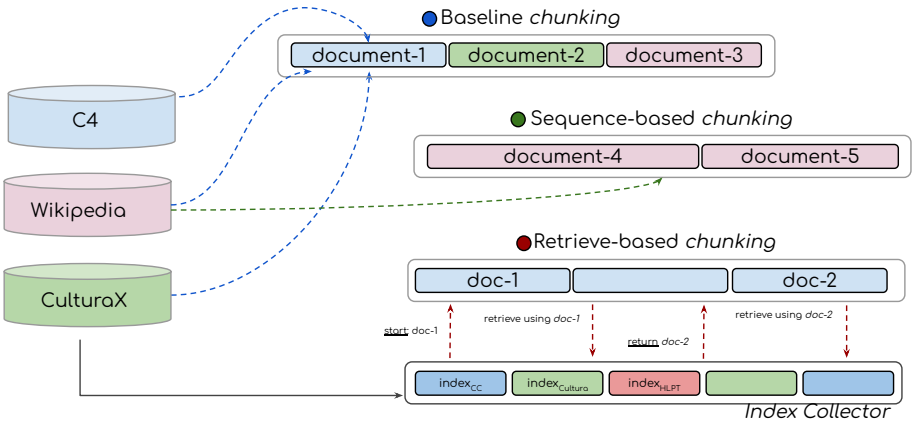

Figure 1: Packing strategies for pre-training chunks construction: *Baseline* randomly samples documents from all corpora to construct pre-training sequences, which can pack documents from different sources; *Sequence-based* randomly samples documents from a single source to construct a sequence; *Retrieve-based* operate via ranking-based construction process. The down block represents a document *Collector* that caches a set of documents randomly sampled between the corpora.

case in different languages as well). As revealed by Zhao et al. [2] as well, this is partly solved by `UniChunk`'s avoidance of packing documents from different distributions, which improves the performance of causal masking models in downstream tasks but still does not allow individual sequences to be selected.

Hence, we use a retrieval-based packing method, which allows operating directly on sequences by improving cross-lingual models' language modeling, in-context learning and generative capabilities by using causal masking and thus paying a small fee for document sorting but achieving tangible results.

Our main findings can be summarised as follows:

- By analyzing different pre-trained strategies in cross-lingual settings we reveal that operating through causal masking and considering the order and patterns sequence represented in documents, leads to significant improvements. In addition, retrieval-based techniques provide resilience and allow for the selection of pre-training sequences by guaranteeing heterogeneity and reducing data (§3).
- We show important benefits on the in-context learning capabilities of downstream models. We observe that in low-resource settings, it is possible to achieve the same performance and in some cases cross-lingual generalisation (in our case, English-Italian) (§4).
- In conclusion, we show that the retrieval-based packing method allowing for a flexible sequence composition process benefits unbalanced cross-lingual learning tangible benefits by using less pre-training data.

2. Pre-Training Strategies

2.1. Packing Approaches

Given $\mathcal{D}_i$ that represents a corpus, and $\mathcal{D} = \bigcup_s \mathcal{D}_s$ denote resulting from the union of such corpora. Specifically, each corpus $\mathcal{D}_s$ is as a set of documents $\mathcal{D}_s = \{d_1, \ldots, d_{|\mathcal{D}_s|}\}$, where each d_i is defined as a sequence of tokens $d_i = \left(x_1, \ldots, x_{|d_i|}\right)$.

The packing strategy involves first selecting a set of documents $\{d_i\}_{i=1}^{n}$ from $\mathcal{D}$, and then packing them into a chunk C with a fixed length $|C| = L$. The documents $\{d_i\}_{i=1}^{n}$ are concatenated by interleaving them with end-of-sentence (`[eos]`) tokens. Hence, C is denoted as:

$$C = \{d_i \oplus \texttt{[eos]} \mid i = 1 \ldots n-1\} \oplus \mathrm{s}(d_n), \quad (1)$$

where [eos] is the end-of-sentence token, s() truncates the last document such that $|C| = L$, and the content of the chunk C is removed from the dataset $\mathcal{D}$ to avoid sampling the same documents multiple times.

Following the strategies proposed in [2], we use three strategies to sample the documents $\{d_i\}_{i=1}^{n}$ from the dataset $\mathcal{D}$ for composing pre-training chunk.

In contrast to the previous works, we use $\alpha \in [0, 1]$ to control the fraction of the corpus used. Hence, we use $\mathcal{S} \subseteq \mathcal{D}$ and $|\mathcal{S}| = \lfloor \alpha \times |\mathcal{D}| \rfloor$.

We define the three strategies (Baseline, Sequence-based and Ranking based) as follow:

Baseline The common baseline approach called `RandomChunk`, with documents $d_i \in \mathcal{D}$ are sampled uni-

formly at random from the entire pre-training corpus $\mathcal{D}$:

$$(\mathcal{D}, \alpha) = \left\{ \bigoplus_{i=1}^{n} d_i \oplus [\text{eos}] \mid d_i \sim \text{Uniform}(\mathcal{S}) \right\} \quad (2)$$

where $\mathcal{S} \subseteq \mathcal{D}$ and $|\mathcal{S}| = \lfloor \alpha \times |\mathcal{D}| \rfloor$. As a result, in `RandomChunk`, a chunk can contain documents from a different source, as shown in Figure 1.

Sequence-based The `UniChunk` approach is sequence-based and respects the sequences of the corpora. Hence, each chunk is composed of documents from a single source corpus $\mathcal{D}_s$:

$$(\mathcal{D}_s, \alpha) = \left\{ \bigoplus_{i=1}^{n} d_i \oplus [\text{eos}] \mid d_i \sim \text{Uniform}(\mathcal{S}_s) \right\} \quad (3)$$

where $\mathcal{S}_s \subseteq \mathcal{D}_s$ and $|\mathcal{S}_s| = \lfloor \alpha \times |\mathcal{D}_s| \rfloor$ and $\mathcal{D}_s \subseteq \mathcal{D}$.

This strategy avoids packing documents from different corpora and allows control over the amount of data utilized from each specific corpus, enhancing efficient usage of computational resources while preserving thematic coherence.

Ranking-based To empower the relevance of documents in pre-training chunks, we use a retriever-based pipeline (BM25-based [4]) to construct pre-training chunks, which we define `Bm25Chunk`. Hence, given a document $d_i \in \mathcal{D}_s$, a sequence of documents $\{d_i\}_{i=1}^{n}$ by $d_{i+1} = \textsc{Retrieve}(d_i, \mathcal{D}_s)$ are retrieved; here, $\textsc{Retrieve}(d_i, \mathcal{D}_s)$ collects the most similar documents to d_i from $\mathcal{D}_s$ using BM25 ranking.

However, since the retrieval process can be computationally heavy due to the size of the pre-training corpus $\mathcal{D}_s$. To improve the efficiency of the retrieval step, a subset $\mathcal{B}_s \subseteq \mathcal{D}_s$ of the corpus $\mathcal{D}_s$ is used, reducing the computational complexity of retrieval as proposed in [2].

In particular, $\mathcal{B}_s \subseteq \mathcal{D}_s$ contains k documents uniformly sampled from $\mathcal{D}_s$. To control the number of utilised documents, we operate via α that regulates the fractions of k. Hence we use $\mathcal{B}_\alpha \subseteq \mathcal{B}_s$ where $|\mathcal{B}_\alpha| = \lfloor \alpha \times |\mathcal{B}_s| \rfloor$.

This approach strategically serves as the retrieval source for constructing pre-training chunks:

$$d_1 \sim \text{Uniform}(\mathcal{B}_s), \quad d_{i+1} = \textsc{Retrieve}(d_i, \mathcal{B}_\alpha).$$

After retrieving a sequence of documents $\{d_i\}_{i=1}^{n}$ from the $\mathcal{B}_\alpha$ for constructing a chunk, the buffer is refilled by sampling novel documents from $\mathcal{D}_s$.

2.2. Masking Approaches

The masking strategy is the other critical stage of language model pre-training, which defines how next-token prediction distributions are conditioned on further tokens in a provided sequence.

Causal Masking In causal masking, each token in a sequence is predicted based on all previous tokens. Specifically, given a chunk $C = (x_1, \ldots, x_{|C|})$, the likelihood of C is given by:

$$P(C) = \prod_{i=1}^{|C|} P(x_i \mid x_1, \ldots, x_{i-1}),$$

where $P(x_i \mid x_1, \ldots, x_{i-1})$ is the probability of the token x_i given previous tokens $x_1, \ldots, x_{i-1}$ in the chunk. During the pre-training, causal masking indicates that, given a chunk C, the likelihood of each token in C is conditioned on all previous tokens, including those that stem from different documents.

Intra-Document Causal Masking In intra-document causal masking, the probability of each token is influenced by the previous tokens within the same document and, consequently, the same context. Hence, using a fraction $\mathcal{S} \subseteq \mathcal{D}$ where $|\mathcal{S}| = \lfloor \alpha \times |\mathcal{D}| \rfloor$ we construct the chunks C asdefined as in §1. The probability of each token d_{ij} belonging to document d_i is only conditioned on the previous tokens within d_i:

$$P(C) = \prod_{i=1}^{n} \prod_{j}^{|d_i|} P\left(d_{ij} \mid d_{i1}, \ldots, d_{i(j-1)}\right), \quad (4)$$

where each d_i is sampled from C as defined above. The models trained using this approach are called `IntraDoc` in the rest of the paper.

3. Language Modeling Settings

Models The implementation is based on the GPT-2 [5]. We pre-train 124 million parameter models using context windows of 256, 512 tokens. To observe the effect of different data compositions, we fix the vocabulary and model parameters described in Appendix A.

Corpora & Settings We combine three high-quality open-source corpora[1] best exemplified from C4, CulturaX, and Wikipedia. We construct the corpus $\mathcal{D}$ by operating through the methods proposed in §2 both on $\mathcal{D}_{En}$ and $\mathcal{D}_{It}$ and then we combine them. Moreover, to observe the impact of the quantity of pre-training instances, we use a scaling factor α that operates during the construction of $\mathcal{D}_{En}$ and $\mathcal{D}_{It}$.

4. Experiments

To analyse the operation of proposed approaches, we evaluate the model perplexities (§4.1), in-context learning (§4.2), understanding (§4.3) and question-answering capabilities (§4.4) under different configurations.

[1]The statistics are reported in Table 4

4.1. Perplexity

We compute the perplexity (PPL) on two different setups: *(i)* models pre-trained with an equal quantity of data and then evaluated on a held-out set of documents where each document is independently treated, *(ii)* models pre-trained with an equal quantity of data scaled by an α factor, which is α in $\{0.1, 0.25, 0.5, 0.75\}$ and then evaluated on a held-out set of documents where each document is independently treated. While the first configuration allows one to observe whether the proposed methods induce overfitting (data-contamination [6]), the second experiment analyses the impact of the amount of data used.

The impact of Sequence Composition Table 1 shows that Bm25Chunk achieves the lowest PPL among the three causal masking models, yielding a lower average PPL compared to RandomChunk (in both settings more than about $+5$) and UniChunk (in both settings around $+3.2$). Increasing the correlation of documents in a sequence empowers the language modelling ability of the pre-trained models. Instead, when considering models trained via intra-document causal masking, it emerges that IntraDoc achieves the lowest PPL compared to the models trained via causal masking.

L	Model	C4	CulturaX	Wiki	Avg.
256	RandomChunk	20.12	19.61	9.89	16.5
	UniChunk	18.83	15.65	8.56	14.3
	Bm25Chunk	**14.96**	15.07	5.23	11.4
	IntraDoc	14.04	**13.57**	**5.08**	**10.7**
512	RandomChunk	19.32	18.76	9.55	15.9
	UniChunk	18.22	15.11	7.89	13.4
	Bm25Chunk	13.85	13.27	5.02	10.7
	IntraDoc	**12.98**	**13.07**	**4.39**	**10.0**

Table 1
Evaluation of perplexity on test set created by sampling the original pre-training corpora (Appendix D). L is the context window for pre-training (next-token accuracy in Appendix B).

Generally, all methods obtain significantly lower PPLs (particularly Bm25Chunk than IntraDoc) in Wikipedia. This phenomenon could imply that the pre-training sources are very common (lower PPL is better-known text), these texts is more influenced by documents with different contexts (misleading contexts) and the proposed strategies can improve this problem.

The role of Quantity Figure 2 shows that Bm25Chunk consistently achieves a lower average PPL than the other approaches even when decreasing the amount of pre-training data. In fact, in both settings (Figure 2), it can be observed that the average PPL of RandomChunk and UniChunk lowers directly as the amount of pre-training data used boosts. While intra-document causal masking performs similarly to Bm25Chunk in resource-based settings (red line and green line Figure 2), improving the intra-document causal masking alpha reduces the PPL less consistently. Finally, it can be observed that Bm25Chunk reaches stable performance even with $\alpha = 0.75$.

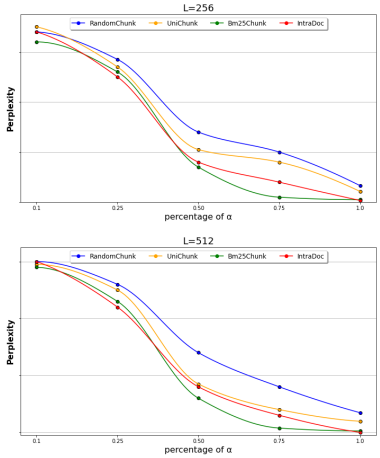

Figure 2: Average Perplexities decreasing training set.

4.2. In-Context Learning

Following Zhao et al. [2], we evaluate the in-context learning abilities of the models using GLUE-X [7] (SST2, CoLA and RTE) both in English and Italian.

Table 2 reports the average in-context learning accuracy values of the models in few-shots settings, using 15 for 256 and 20 demonstrations for the 512 model, respectively. Bm25Chunk yields a higher average accuracy than RandomChunk for 256 ($+5.12\%$) and 512 ($+1.55\%$). These demonstrate that increasing the correlation of the documents in pre-training chunks improves the models' in-context learning abilities.

Figure 3, we report the average accuracy using different numbers of few-shot demonstrations. Bm25Chunk has an on-par accuracy with IntraDoc on the 256 setting; however, IntraDoc obtains a significantly higher accuracy than Bm25Chunk on the 512 setting. Finally, RandomChunk and UniChunk obtain comparable results using different context lengths, and they do not consistently improve accuracy when increasing the number of demonstrations. This might be due to the tighter levels of distraction in both settings, which use arbitrary packing strategies.

L	Model	SST2	CoLA	RTE	Avg.
256	RandomChunk	50.53	60.62	24.76	45.33
	UniChunk	56.13	62.68	18.73	45.72
	Bm25Chunk	**62.12**	**64.06**	**25.16**	**50.45**
	IntraDoc	53.22	61.16	24.23	46.20
512	RandomChunk	55.13	62.85	36.38	51.38
	UniChunk	**58.53**	63.04	22.12	47.85
	Bm25Chunk	60.30	63.21	35.26	52.93
	IntraDoc	59.32	**65.62**	**36.65**	**53.81**

Table 2
Average In-context learning performance evaluated by text classification accuracy across three tasks. Accuracies for English and Italian are reported in Appendix E.

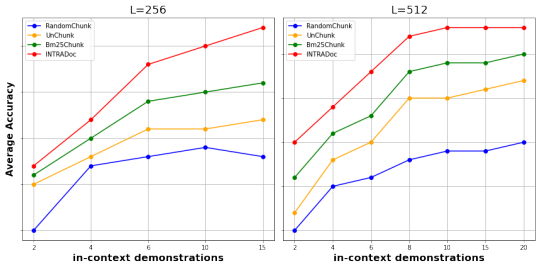

Figure 3: Average in-context learning accuracy using different numbers of input demonstrations.

L	Model	MLQA	XCOPA	SQuAD	Avg.
256	RandomChunk	21.48	30.21	**28.04**	26.5
	UniChunk	23.97	32.19	27.16	27.7
	Bm25Chunk	28.18	33.97	27.26	29.8
	IntraDoc	**33.63**	**38.05**	**30.51**	**34.0**
512	RandomChunk	26.05	31.93	31.39	29.7
	UniChunk	27.14	33.34	31.22	30.5
	Bm25Chunk	30.71	35.82	34.85	33.7
	IntraDoc	**32.42**	**37.71**	**36.04**	**35.2**

Table 3
Evaluation results of natural language understanding, commonsense reasoning and QA tasks.

4.3. Understanding & Commonsense

We evaluate the pre-trained models on natural language understanding, commonsense reasoning tasks (i.e., XSQuAD [8], XCOPA [9]), and question-answering (i.e., MLQA [10]). It emerges that Bm25Chunk outperforms RandomChunk and UniChunk in all tasks, confirming that increasing the similarity of documents in pre-training chunks improve understanding abilities. Specifically, Bm25Chunk obtains a significantly better accuracy on MLQA, showing it can operate in-context information provided in the input question.

However, even though Bm25Chunk archives solid performances, IntraDoc obtains the best average performance. It indicates that eliminating potential distractions from unrelated documents and learning each document separately empowers understanding and generation abilities. This finding is different from the ideas in previous works, which suggested that pre-training with multiple documents in one context and adding distraction in context during pre-training benefit in-context and understanding ability.

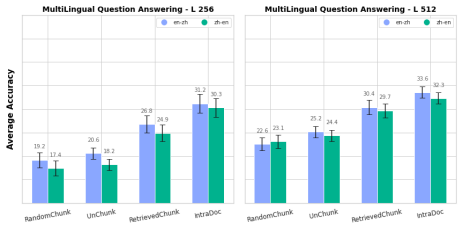

Figure 4: Evaluation results of MultiLingual Question Answering by providing cross-lingual input (en-it means context in English and question in Italian and vice versa as described in Appendix C).

4.4. Multilinguality

To assess code-switching abilities, we experimented with cross-lingual input by operating with MLQA. We crossed the languages, delivering contexts in English and questions in Italian and vice versa (Appendix C). Figure 4 show that Bm25Chunk outperforms both RandomChunk and intra-document causal masking. At the same time, IntraDoc, as discussed in §4.3 for MLQA, outperforms Bm25Chunk. This result confirms that IntraDoc's performance is not only related to monolingual learning sequences but also more complex dynamics.

5. Conclusion

The role of pre-training sampling is a strategic component. We analyse the impact of sequencing by pre-training several language models on multilingual corpora. We showed that causal masking involves misleading documents that confound the pre-training of language models and impact the performance in downstream tasks. Hence, we find that improving sequence correlation in pre-training chunks reduces potential distractions while improving the performance of language models without reducing pre-training efficiency. In the future, we will study whether these findings archive benefits in fine-tuning pipelines [11, 12, 13, 14, 15, 16] as well.

References

[1] L. Ranaldi, G. Pucci, F. M. Zanzotto, Modeling easiness for training transformers with curriculum learning, in: R. Mitkov, G. Angelova (Eds.), Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing, INCOMA Ltd., Shoumen, Bulgaria, Varna, Bulgaria, 2023, pp. 937–948. URL: https://aclanthology.org/2023.ranlp-1.101.

[2] Y. Zhao, Y. Qu, K. Staniszewski, S. Tworkowski, W. Liu, P. Miłoś, Y. Wu, P. Minervini, Analysing the impact of sequence composition on language model pre-training, 2024. URL: https://arxiv.org/abs/2402.13991. arXiv:2402.13991.

[3] W. Shi, S. Min, M. Lomeli, C. Zhou, M. Li, V. Lin, N. A. Smith, L. Zettlemoyer, S. Yih, M. Lewis, In-context pretraining: Language modeling beyond document boundaries, ArXiv abs/2310.10638 (2023). URL: https://api.semanticscholar.org/CorpusID:264172290.

[4] S. Robertson, H. Zaragoza, The probabilistic relevance framework: Bm25 and beyond, Found. Trends Inf. Retr. 3 (2009) 333–389. URL: https://doi.org/10.1561/1500000019. doi:10.1561/1500000019.

[5] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. 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. 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: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., 2020, pp. 1877–1901. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[6] F. Ranaldi, E. S. Ruzzetti, D. Onorati, L. Ranaldi, C. Giannone, A. Favalli, R. Romagnoli, F. M. Zanzotto, Investigating the impact of data contamination of large language models in text-to-SQL translation, in: L.-W. Ku, A. Martins, V. Srikumar (Eds.), Findings of the Association for Computational Linguistics ACL 2024, Association for Computational Linguistics, Bangkok, Thailand and virtual meeting, 2024, pp. 13909–13920. URL: https://aclanthology.org/2024.findings-acl.827. doi:10.18653/v1/2024.findings-acl.827.

[7] L. Yang, S. Zhang, L. Qin, Y. Li, Y. Wang, H. Liu, J. Wang, X. Xie, Y. Zhang, GLUE-X: Evaluating natural language understanding models from an out-of-distribution generalization perspective, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Findings of the Association for Computational Linguistics: ACL 2023, Association for Computational Linguistics, Toronto, Canada, 2023, pp. 12731–12750. URL: https://aclanthology.org/2023.findings-acl.806. doi:10.18653/v1/2023.findings-acl.806.

[8] P. Rajpurkar, J. Zhang, K. Lopyrev, P. Liang, Squad: 100, 000+ questions for machine comprehension of text, in: J. Su, X. Carreras, K. Duh (Eds.), Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016, The Association for Computational Linguistics, 2016, pp. 2383–2392. URL: https://doi.org/10.18653/v1/d16-1264. doi:10.18653/V1/D16-1264.

[9] E. M. Ponti, G. Glavaš, O. Majewska, Q. Liu, I. Vulić, A. Korhonen, XCOPA: A multilingual dataset for causal commonsense reasoning, in: B. Webber, T. Cohn, Y. He, Y. Liu (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 2362–2376. URL: https://aclanthology.org/2020.emnlp-main.185. doi:10.18653/v1/2020.emnlp-main.185.

[10] P. Lewis, B. Oguz, R. Rinott, S. Riedel, H. Schwenk, MLQA: Evaluating cross-lingual extractive question answering, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 7315–7330. URL: https://aclanthology.org/2020.acl-main.653. doi:10.18653/v1/2020.acl-main.653.

[11] L. Ranaldi, G. Pucci, Knowing knowledge: Epistemological study of knowledge in transformers, Applied Sciences 13 (2023). URL: https://www.mdpi.com/2076-3417/13/2/677. doi:10.3390/app13020677.

[12] L. Ranaldi, G. Pucci, Does the English matter? elicit cross-lingual abilities of large language models, in: D. Ataman (Ed.), Proceedings of the 3rd Workshop on Multi-lingual Representation Learning (MRL), Association for Computational Linguistics, Singapore, 2023, pp. 173–183. URL: https://aclanthology.org/2023.mrl-1.14. doi:10.18653/v1/2023.mrl-1.14.

[13] L. Ranaldi, G. Pucci, F. Ranaldi, E. S. Ruzzetti, F. M. Zanzotto, A tree-of-thoughts to broaden multi-step reasoning across languages, in: K. Duh, H. Gomez, S. Bethard (Eds.), Findings of the Association for Computational Linguistics: NAACL 2024, Association for Computational Linguistics, Mexico City, Mexico, 2024, pp. 1229–1241. URL: https:

//aclanthology.org/2024.findings-naacl.78. doi:10.18653/v1/2024.findings-naacl.78.

[14] L. Ranaldi, G. Pucci, A. Freitas, Does the language matter? curriculum learning over neo-Latin 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. 5212–5220. URL: https://aclanthology.org/2024.lrec-main.464.

[15] L. Ranaldi, A. Freitas, Aligning large and small language models via chain-of-thought reasoning, in: Y. Graham, M. Purver (Eds.), Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, St. Julian's, Malta, 2024, pp. 1812–1827. URL: https://aclanthology.org/2024.eacl-long.109.

[16] L. Ranaldi, A. Freitas, Self-refine instruction-tuning for aligning reasoning in language models, in: Y. Al-Onaizan, M. Bansal, Y.-N. Chen (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 2325–2347. URL: https://aclanthology.org/2024.emnlp-main.139.

A. Pre-training Corpora

In our experiments, we use the GPT-2 small, the 124 million model with 12 layers, a hidden size of 768, and 12 attention heads. We use a batch size of 0.5 million tokens for both the models with 256 and 512 context window sizes and pre-train models using 20B tokens with 100,000 steps. We use Adam optimiser with $\beta_1 = 0.90$, $\beta_2 = 0.95$, a weight decay of 0.1, and a cosine learning rate scheduler. The peak learning rate is 3×10^{-4}, decreasing to 3×10^{-5} at the end. We perform the experiments using 16 Nvidia RTX A6000 with 48GB of VRAM.

Subset	# documents	# words
C4 (it)	$\sim 8M$	$\sim 4B$
CulturaX (it)	$\sim 2.5M$	$\sim 2.6M$
Wikipedia (it)	$\sim 1.5M$	$\sim 780M$
C4 (it)	$\sim 8M$	$\sim 3.4B$
CulturaX (it)	$\sim 2.5M$	$\sim 2.1M$
Wikipedia (it)	$\sim 1.5M$	$\sim 760M$

Table 4
Size of pre-training corpora. For computational reasons, we produced equivalent samples for both English and Italian.

B. Next Token Accuracy of Pre-Trained Language Models

In addition to PPL, we report the next token accuracy of pre-trained language models in Table 5.
The "next-token accuracy" is calculated as follows:
Specifically we define Acc as:

$$\text{Acc} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(\hat{y}_i = y_i) \tag{5}$$

where:

- N is the total number of tokens in the test set.
- $\hat{y}_i$ is the token predicted by the model at position i.
- y_i is the correct (ground truth) token at position i.
- $\mathbb{I}$ is the indicator function, which is 1 if $\hat{y}_i = y_i$ and 0 otherwise.

L	Model	C4	CulturaX	Wikipedia	Avg.
256	RandomChunk	0.242	0.431	0.336	0.336
	UniChunk	0.248	0.463	0.415	0.375
	Bm25Chunk	0.332	0.451	0.424	0.402
	IntraDoc	**0.357**	**0.472**	**0.442**	**0.423**
512	RandomChunk	0.346	0.456	0.368	0.393
	UniChunk	0.389	0.462	0.405	0.419
	Bm25Chunk	0.419	0.493	0.423	0.445
	IntraDoc	**0.440**	**0.498**	**0.463**	**0.467**

Table 5
Evaluation of next token accuracy on proposed test-set.

C. Multilingual Question Answering Examples

Lang	Context	Question	Target Answer
en	Barack Obama was the 44th President of the United States, serving two terms from 2009 to 2017.	Who was the 44th President of the United States?	Barack Obama
it	Barack Obama è stato il 44° Presidente degli Stati Uniti, in carica per due mandati dal 2009 al 2017.	Chi è stato il 44° Presidente degli Stati Uniti?	Barack Obama
en-it	Barack Obama was the 44th President of the United States, serving two terms from 2009 to 2017.	*Chi è stato il 44° Presidente degli Stati Uniti?*	Barack Obama
it-en	Barack Obama è stato il 44° Presidente degli Stati Uniti, in carica per due mandati dal 2009 al 2017.	*Who was the 44th President of the United States?*	Barack Obama

Table 6
Examples from the MLQA dataset in **English**, **Italian** and **Cross-lingual**.

D. In-context Learning performances English and Italian

This section reports the results obtained on the tasks introduced in Section 4.2. To conduct a more detailed analysis, we have used the original (English) and Italian versions of three tasks belonging to the GLUE family. We selected SST2, CoLA, and RTE. The bilingual versions were taken from the contribution previously proposed by Yang et al. [7].

	Model	SST2-En	CoLA-En	RTE-En	Avg.
256	RandomChunk	51.34	61.73	25.71	46.26
	UniChunk	57.16	63.21	19.17	43.15
	Bm25Chunk	**61.9**	**65.02**	**26.31**	50.42
	IntraDoc	53.39	61.67	25.27	46.76
512	RandomChunk	55.49	63.42	38.19	52.46
	UniChunk	59.16	63.12	21.87	48.02
	Bm25Chunk	**60.81**	64.69	36.23	53.93
	IntraDoc	59.21	**66.25**	**36.19**	53.73

Table 7
In-context learning performance evaluated by text classification accuracy across three **English** tasks.

	Model	SST2-It	CoLA-It	RTE-It	Avg.
256	RandomChunk	49.41	59.62	23.51	44.17
	UniChunk	55.13	62.92	18.32	46.76
	Bm25Chunk	**61.24**	**63.07**	**23.92**	49.40
	IntraDoc	52.93	60.81	23.92	46.08
512	RandomChunk	54.71	62.63	34.36	50.64
	UniChunk	57.92	62.94	22.46	47.82
	Bm25Chunk	**59.83**	63.38	34.25	52.36
	IntraDoc	59.06	**65.23**	**35.16**	52.55

Table 8
In-context learning performance evaluated by text classification accuracy across three **Italian** tasks.

E. Understanding and Commonsense performances English and Italian

This section reports the results obtained on the tasks introduced in Section 4.3. We have used the original (English) and Italian versions of MLQA, XCOPA, and SQuAD to conduct a more detailed analysis.

L	Model	MLQA	XCOPA	SQuAD	Avg.
256	RandomChunk	22.63	30.71	30.52	30.22
	UniChunk	24.09	23.15	27.34	24.83
	Bm25Chunk	29.16	34.19	27.16	30.11
	IntraDoc	**34.06**	**38.21**	**30.85**	34.3
512	RandomChunk	26.63	32.16	31.82	30.32
	UniChunk	27.05	33.26	31.54	30.65
	Bm25Chunk	30.66	36.51	34.73	34.08
	IntraDoc	**32.88**	**38.15**	**38.23**	36.23

Table 9
Evaluation results of natural language understanding, commonsense reasoning and QA tasks in **English**.

L	Model	MLQA	XCOPA	SQuAD	Avg.
256	RandomChunk	20.33	29.62	30.18	29.31
	UniChunk	23.85	23.42	26.73	25.06
	Bm25Chunk	27.21	33.16	27.32	29.05
	IntraDoc	**33.26**	**37.88**	**30.18**	33.65
512	RandomChunk	25.88	31.78	30.97	x.x
	UniChunk	27.23	33.42	30.94	30.32
	Bm25Chunk	30.77	35.92	34.66	33.42
	IntraDoc	**31.97**	**37.28**	**38.46**	35.64

Table 10
Evaluation results of natural language understanding, commonsense reasoning and QA tasks in **Italian**.