BABILong-ITA: a new benchmark for testing Large Language Models effective context length and a Context Extension Method

Fabio Tamburini[1,*]

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

Abstract

This paper introduces a new benchmark designed to evaluate the effective context length handled by Large Language Models (LLMs) in Italian. Following the structure of the five core tasks from the English BABILong dataset, we created an equivalent benchmark tailored for Italian. We used it to assess the context management capabilities of several prominent LLMs, both small and large, pretrained from scratch or fine-tuned specifically for Italian. Additionally, we tested a context extension technique called "SelfExtend" that does not require any training or fine-tuning phase, measuring its effectiveness using our proposed benchmark.

Keywords

Large Language Models, context length evaluation, new benchmark, Italian, context extension

1. Introduction

As the capabilities of Large Language Models (LLMs) continue to advance, one of the most critical areas of improvement lies in their ability to process and retain information over extended sequences of text, a feature commonly referred to as context length. Traditional benchmarks for evaluating LLMs focus on accuracy, reasoning, and generation quality, but often overlook systematic assessment of how well a model can operate when presented with extremely long input sequences.

LLMs long context is crucial for Retrieval-Augmented Generation (RAG) because it allows the model to process and reason over more retrieved information at once. In RAG systems, external documents or chunks of text are retrieved based on a query and then passed to the LLM to generate accurate and contextually relevant answers. A longer context window means the model can consider more documents or larger portions of documents simultaneously, reducing the need to truncate or summarise input data. This leads to better comprehension, improved factual accuracy, and more coherent responses, especially for complex or multi-part queries.

Evaluating the context length capabilities of LLMs is crucial for understanding their practical utility in real-world applications requiring long-range reasoning, document understanding, and multi-turn conversations. Over the past years, several standardised benchmarks have been developed to assess and compare the performance of LLMs across varying context lengths.

A widely cited benchmark framework is the Kamradt's 'Needle-in-a-Haystack'[1] which probes a model's ability to retrieve a small piece of relevant information embedded in a long, distractor-filled sequence. This test is considered a litmus test for whether models truly attend to long-range dependencies rather than relying on heuristics or recency biases.

Another critical benchmark is 'Passage Retrieval and Question Answering' over long contexts, exemplified by datasets such as 'NarrativeQA' [1] and 'HotpotQA' [2]. These datasets require models to maintain coherence and extract pertinent information across several paragraphs or documents. The 'BookSum' benchmark [3] further extends this approach by evaluating abstractive summarisation over entire books, posing an extreme challenge to context handling.

To assess performance on computationally efficient long-context processing, the 'Long Range Arena' provides a suite of tasks including image classification, text retrieval, and list sorting, adapted to sequence modelling tasks with sequences ranging from 1k to 16k tokens [4]. While not all tasks are purely devoted to natural language processing, they benchmark architectural innovations like sparse attention and memory-efficient transformers.

'LongBench' [5] provides comprehensive testbeds across domains covering key long-text application areas including single-doc QA, multi-doc QA, summarisation, few-shot learning, synthetic tasks, and code completion in both English and Chinese, evaluating both performance scaling and fidelity to far-positioned inputs.

An et al. [6] present a new evaluation suite 'L-Eval'

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)

[1]https://github.com/gkamradt/LLMTest_NeedleInAHaystack.git

containing 20 sub-tasks, 508 long documents, and more than 2,000 human-labelled query-response pairs including diverse task types, domains, and input lengths.

Taken together, these benchmarks form a multi-faceted suite of tools that not only test LLMs for maximum supported context length but also probe their effective use of context. As models scale to handle millions of tokens, developing robust and generalisable long-context benchmarks remains an active area of research, especially for languages different from English.

Regarding the techniques for increasing context 'awareness' in transformers, recent works have introduced scaling techniques specifically targeting context length extrapolation. For example, Press et al. [7] proposed the in-Context Learning Extrapolation to test model performance when context lengths at inference time far exceed those seen during training. Considering this, we could refer to a recent interesting survey on techniques for extending transformers context by Wang et al. [8].

Another English benchmark, relevant to this work, is 'BABILong' [9], a benchmark specifically designed to evaluate the maximum usable context length of large language models. BABILong provides a controlled and extensible framework for measuring how effectively LLMs can retrieve and use information embedded at various positions within long input contexts. The benchmark simulates real-world scenarios where crucial information may appear early in a document and must be recalled accurately much later, such as in code completion, document summarisation, and legal or scientific reasoning tasks. Each BABILong instance presents the model with a structured sequence containing query-relevant and distractor content spread over thousands to potentially millions of tokens. The model is then tasked with answering queries or completing sequences that require precise recollection of target information, making it possible to assess the degradation of performance as a function of input length.

Unlike traditional evaluations, BABILong systematically varies the distance between the query and its corresponding reference information, enabling granular analysis of context window utilisation and scaling properties across different architectures. The benchmark supports plug-and-play integration with both decoder-only and encoder-decoder models, and it is agnostic to pretraining data, making it suitable for comparative studies across proprietary and open-source models.

In summary, BABILong provides a scalable, interpretable, and model-agnostic benchmark for long-context reasoning and memory fidelity and it is a very useful tool for researchers and practitioners seeking to push the boundaries of efficient long-sequence modelling in large-scale language systems. Moreover, it can be easily extended to other languages: the goal of this work regards the extension of BABILong to Italian, allowing for a careful testing and benchmarking of LLMs that natively handle the Italian language.

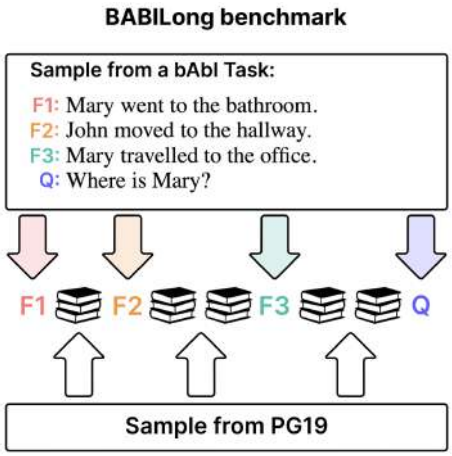

Figure 1: BABILong schema for generating tasks: task facts are hidden into distractor text fragments extracted from PG19 (picture from [9]).

2. A new benchmark for Italian

BABILong extends the bAbI benchmark [10], which consists of 20 tasks designed to evaluate basic aspects of reasoning. These tasks are generated by simulating interactions among characters and objects across various locations, each represented as a fact, such as "Mary traveled to the office." The challenge is to answer questions based on the facts generated in the current simulation, such as "Where is Mary?" The tasks in bAbI vary in the number of facts, question complexity, and the reasoning skills they assess, including spatial and temporal reasoning, deduction, and coreference resolution.

Solving tasks that require long-context processing demands that a model effectively identify and attend to relevant information embedded within extensive irrelevant content. To emulate this scenario, they embed the core task sentences within passages of distractor text sampled from a closely related distribution (see Figure 1). Each example is constructed by progressively appending sentences from the background corpus, preserving their natural order, until the desired total length is achieved. This approach decouples the evaluation context length from the intrinsic length of the original task, thereby enabling the assessment of models capable of handling inputs extending to millions of tokens. As background material, they used books from the PG19 dataset [11], chosen for their substantial length and naturally occurring long-form narrative structure.

We reproduced the same process proposed in BABI-

Long by, first, translating English sentences belonging to BABILong tasks leveraging Google Translate and then using the Project Gutemberg[2] (PG) Italian free texts as base corpus for extracting distractor fragments.

Given that all the major evaluations in the BABILong paper [9] were performed considering only the first five tasks, namely QA1-QA5, we decided to translate and post-process only these five tasks and insert them into BABILong-ITA.

In order to build a reliable and effective Italian benchmark we had to manually revise and adapt automatic translations ensuring a good adherence to common Italian language adjusting translation artifacts or wrong translations. In particular, we had to manage these phenomena:

- **Proper Names translation**: Google Translate did not translate English proper names of people involved in the task, thus we have to replace them consistently with common Italian proper names, e.g. 'John'->'*Giovanni*', 'Mary'->'*Maria*', etc.
- **Object/Place Simplification**: the automatic translation tended, in some cases, to translate single English words into Italian multiword expressions artificially increasing tasks difficulty. We simplify objects/places translations like 'bedroom'->'*camera da letto*'->'*camera*' and 'football'->'*pallone da calcio*'->'*pallone*', etc.
- **Verb Tenses**: for expressing past events English consistently use the past tense while in Italian, even if the equivalent past tense '*passato remoto*' is grammatically correct, is much more common using the '*passato prossimo*'. We then adapted the translations replacing all these tenses, e.g. '*andò*'->'*è andato/a*', '*posò*'->'*ha posato*' and '*si spostò*->'*si è spostato/a*' adapting the suffixes to the sentence subject preserving the correct grammatical agreement.
- **Proposition Correction**: sometimes Google Translate generates inappropriate translations from the point of view of the used prepositions; we corrected them, for example '*John si recò al giardino*'->'*Giovanni si è recato in giardino.*' or '*Mary andò nel corridoio*'->'*Maria è andata in corridoio*', ensuring a better adherence to the most common use of them.
- **Translation Mistake Corrections**: sometimes, especially when translating questions with implicit referents, Google Translate rendered incorrect Italian sentences that we have to carefully check and correct also by leveraging regular expressions: for example 'What is the kitchen west of?'->'*Qual è la cucina a ovest?*'->'*La cucina è a ovest di che cosa?*'.

While we could have incorporated a broader range of state/position-changing predicates in the translations, we chose to adhere to the original selections, as the English benchmark did not include such variations.

Table 1 shows one example for each BABILong-ITA task without the insertions of any distractor texts (0k configuration).

3. Benchmark evaluation

In order to test the effectiveness of the new proposed benchmark and to grasp some idea about the performance of the most relevant models able to effectively handle the Italian language, we performed a set of experiments involving quite a large set of LLMs.

First of all, we considered the new models presented in 2024 and trained from scratch on Italian: the first by the SapianzaNLP group[3], namely *sapienzanlp/Minerva-7B-base-v1.0* and *sapienzanlp/Minerva-7B-instruct-v1.0*, and, second, the largest model proposed by iGenius/CINECA using the unofficial conversion *sapienzanlp/modello-italia-9b-bf16* for simplicity. We considered also two fine-tuned model from DeepMount00, namely *DeepMount00/Qwen2-1.5B-Ita* and *DeepMount00/Mistral-Ita-7b*, a model from Microsoft, *microsoft/Phi-4-mini-instruct*, one from meta, *meta-llama/Llama-3.1-8B-Instruct* both in its original and quantised form relying on *bartowski/Meta-Llama-3.1-8B-Instruct-Q4_K_S* and, finally, two models from Google, *google/gemma-3-4b-it* and the huge *google/gemini-2.0-flash*. All models were downloaded from the HuggingFace model repository[4] and used on a local server except for *gemini-2.0-flash* that was queried using the Google API.

3.1. Experiments setting

In BABILong, the authors consider performance satisfactory if the accuracy of an answer exceeds 85% and a complete failure if it is below 30%. Of course, as the authors said, this definition of "satisfactory performance" is not universal and should be adapted to the specific task at hand.

The comparison with the correct result follows the original BABILong evaluation method: the LLM output is lowercased, and the first valid target it names is considered as the LLM answer and compared with the gold target in order to compute model accuracy.

[2] https://www.gutenberg.org/

[3] https://nlp.uniroma1.it/minerva/

[4] https://huggingface.co/

QA1 single-supporting-fact
Context: *Sandra si è diretta verso la cucina. Daniele si è diretto verso il bagno. Maria è andata in giardino.* ***Maria si è recata in ufficio****. Sandra si è recata in camera. Giovanni si è recato in ufficio. Sandra si è recata in ufficio. Sandra si è trasferita in cucina.*
Question: *Dov'è Maria?* Answer: ***ufficio****.*
QA2 two-supporting-facts
Context: *Sandra si è diretta verso il corridoio. Giovanni si è diretto verso il bagno. Sandra ha afferrato il pallone lì. Daniele si è recato in camera. Giovanni ha preso il latte lì. Giovanni ha lasciato cadere il latte. Sandra si è trasferita in giardino. Daniele è tornato in corridoio. Sandra ha buttato via il pallone. Giovanni si è spostato in corridoio. Giovanni è tornato in giardino. Sandra è andata in cucina. Daniele si è trasferito in camera. Sandra si è diretta verso il corridoio. Sandra si è trasferita in cucina. Giovanni si è recato in ufficio.* ***Sandra è andata in giardino. Sandra ha afferrato il pallone lì. Sandra ha posato lì il pallone****. Daniele è tornato in cucina.*
Question: *Dov'è il pallone?* Answer: ***giardino****.*
QA3 three-supporting-facts
Context: *Maria è andata in ufficio. Sandra si è spostata in corridoio. Sandra ha afferrato il pallone. Maria ha preso lì la mela. Sandra si è recata in giardino. Daniele si è spostato in corridoio. Sandra ha posato il pallone. Daniele è andato in camera. Sandra ha preso il pallone. Maria ha posato la mela. Maria è tornata in bagno. Giovanni si è spostato in bagno. Giovanni è andato in corridoio. Sandra ha posato il pallone. Daniele si è diretto verso il corridoio. Sandra ha raccolto il pallone. Sandra si è recata in ufficio. Daniele si è recato in bagno. Daniele è tornato in ufficio. Daniele si è recato in cucina. Sandra ha raccolto la mela lì. Sandra ha buttato lì la mela. Sandra ha lasciato cadere il pallone. Giovanni si è recato in giardino. Maria si è recata in giardino. Sandra ha afferrato il pallone lì. Sandra ha buttato lì il pallone. Sandra si è diretta verso la cucina. Maria si è trasferita in camera. Maria è andata in corridoio. Sandra si è diretta verso il corridoio. Giovanni è andato in cucina. Sandra si è recata in bagno. Daniele è tornato in bagno. Giovanni si è trasferito in ufficio. Giovanni ha preso il latte. Giovanni si è diretto verso il bagno. Daniele è tornato in camera. Maria si è recata in camera. Daniele si è diretto verso il corridoio. Giovanni si è trasferito in camera. Sandra si è recata in giardino. Daniele è tornato in cucina. Giovanni ha lasciato il latte. Daniele si è recato in ufficio. Daniele ha preso il pallone. Maria è andata in corridoio.* ***Daniele ha afferrato la mela lì****. Giovanni si è diretto verso il bagno. Giovanni si è diretto verso il corridoio. Giovanni è andato in ufficio. Giovanni è tornato in cucina. Maria si è recata in ufficio.* ***Daniele è tornato in giardino. Daniele è andato in camera. Daniele si è spostato in bagno. Daniele è tornato in giardino****. Sandra è tornata in bagno.* ***Daniele è andato in camera. Daniele ha lasciato la mela****. Daniele ha lasciato il pallone. Daniele ha afferrato il pallone.*
Question: *Dov'era la mela prima di essere in camera?* Answer: ***giardino****.*
QA4 two-arg-relations
Context: ***Il giardino si trova a ovest della camera. L'ufficio si trova a est della camera.***
Question: *La camera è a est di che cosa?* Answer: ***giardino****.*
QA5 three-arg-relations
Context: *Enrico ha preso il pallone lì. Enrico si è recato in giardino.* ***Enrico ha passato il pallone a Giovanni****. Maria è andata in cucina.* ***Giovanni ha passato il pallone a Enrico. Enrico ha consegnato il pallone a Giovanni****. Maria ha preso il latte lì. Giovanni si è diretto verso la cucina. Giovanni si è trasferito in giardino. Daniele si è recato in camera.*
Question: *Chi ha ricevuto il pallone?* Answer: ***Giovanni****.*

Table 1
BABILong-ITA examples. This table shows the "0k" configuration without distracting text. In the longer context configurations (1k, 2k, 4k...) fragment of texts from PG have been inserted between context sentences as distractors following the original BABILong schema.

3.2. Results

Figure 2 presents the average retrieval accuracy across all tasks for each evaluated LLM on the BABILong-ITA benchmark.

LLMs trained from scratch in Italian, specifically Minerva and Modello-Italia, generally show low performance. However, within their maximum supported context length of 4k tokens, their performance remains relatively stable across all tested lengths.

The fine-tuned models from DeepMount00 demonstrate consistently poor retrieval performance. Despite a declared maximum context length of 32k tokens, they struggle significantly even at much shorter lengths. Similar observations apply to Phi-4, which fails to achieve satisfactory results even at just 1/16 of its maximum declared context window.

Google's Gemma3 shows slightly better performance, managing to handle contexts up to approximately 1/8 of its maximum declared length. Conversely, Gemini-2.0-flash, with a nominal maximum context length of 1 million tokens, solves fewer than 50% of the tasks at 128k, an underwhelming result given its scale.

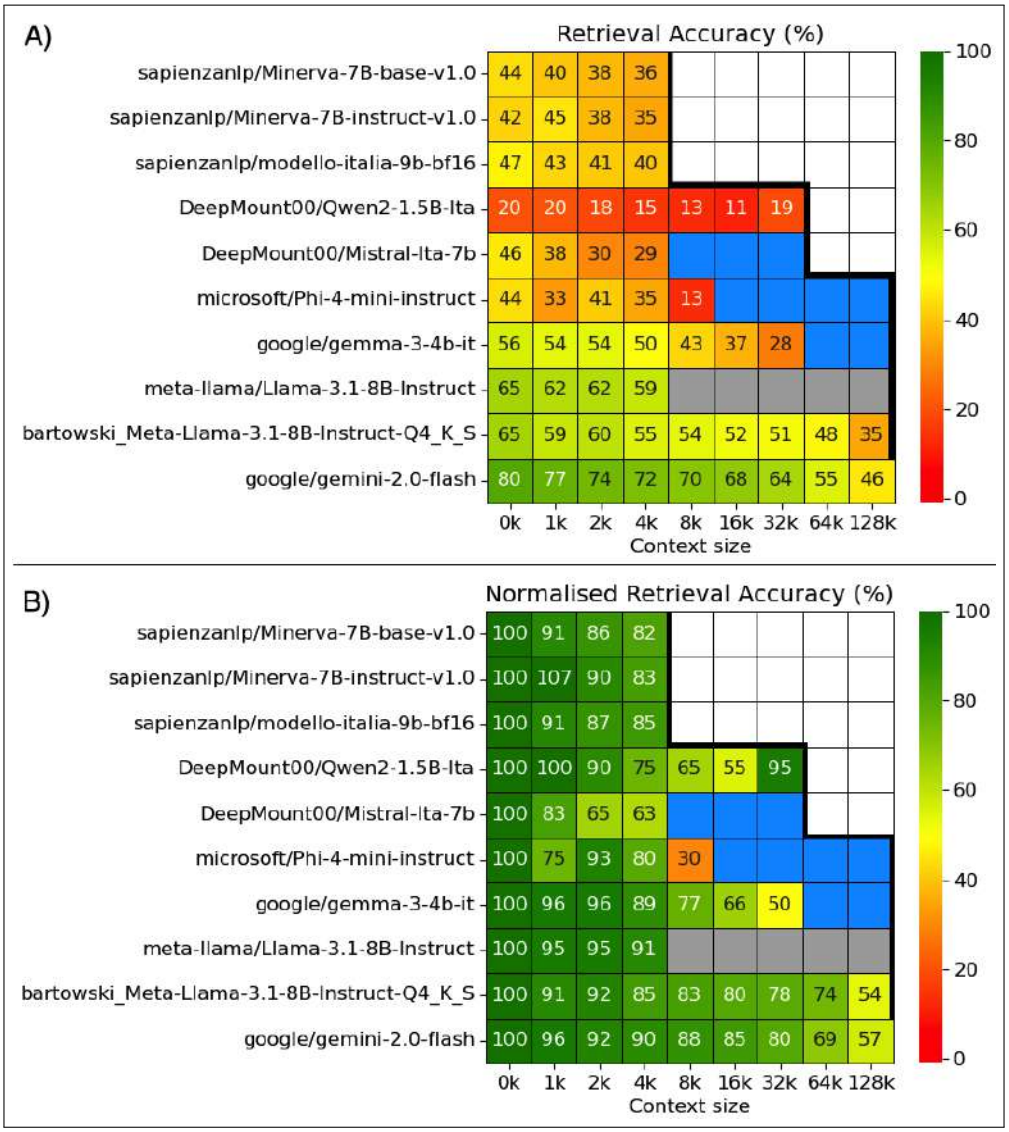

Figure 2: BABILong-ITA evaluation results for tested LLMs averaged over the five proposed tasks. The thick black line marks the maximum context length for a given model (for Gemini-2.0-flash the official limit is 1000k tokens). Blue squares indicate cells not computed for computational restriction reasons, but actually not useful for the evaluation because smaller contexts already presented very low retrieval accuracies. On the contrary, gray cells marks combinations that we were not able to calculate that are approximated by the corresponding quantised model. In A) we have original accuracies, while in B) we normalised the accuracy w.r.t. the "0k" case to show the relative reduction of accuracy obtained increasing the test context size.

Among the tested models, LLaMA-3.1-8B stands out as the most effective. Although we completely evaluated only its quantised version, which performs slightly below the full model, it successfully retrieves 35% of the hidden information even at the maximum declared context length. It appears to offer an excellent balance between local deployment feasibility and performance, trailing only slightly behind the much larger Gemini-2 model.

Figure 3 presents the per-task performance of the two best-performing LLMs tested, namely Gemini-2.0-flash and the quantised version of LLaMA-3.1-8B. The QA2 and QA3 tasks are notably more complex than the others, with both models struggling to retrieve the target information in QA3, even within very short contexts.

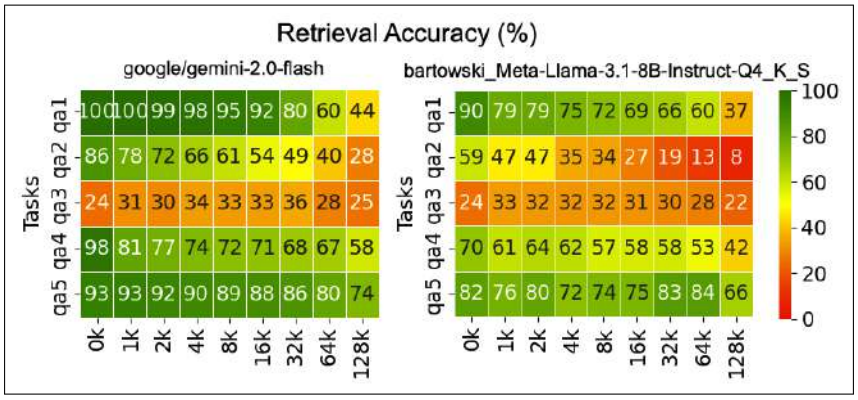

Figure 3: Per task BABILong-ITA evaluation results for the best two tested LLMs.

Given these results and the smooth transitions across different context lengths, we can conclude that BABILong-ITA appears to be a reliable benchmark for testing the effective context length of LLMs.

4. Extending Large Language Models Context Length

Extending the context length of LLMs is a key research direction aimed at improving their ability to reason over long documents, maintain dialogue coherence, and process extensive sequences of information.

Several approaches have emerged to address the computational and architectural challenges associated with long-context modeling:

- **Sparse Attention and Efficient Transformers**. One class of techniques involves modifying the attention mechanism to reduce its quadratic complexity with respect to sequence length. Models such as *Longformer* [12], *BigBird* [13], and *Reformer* [14] introduce sparse or locality-sensitive hashing attention patterns to enable efficient processing of longer sequences. These methods trade off some global attention capacity for linear or sub-quadratic scaling, allowing context lengths up to tens of thousands of tokens.
- **Memory-Augmented Models**. These models incorporate external memory buffers to persist information across long sequences. *Transformer-XL* [15] uses a segment-level recurrence mechanism, enabling longer context windows by caching hidden states across segments. Similarly, models like *Compressive Transformer* [11] compress and store previous activations to extend memory capacity while maintaining computational tractability.
- **Position Encoding Innovations**. Absolute positional encodings pose a limitation on extrapolation beyond trained sequence lengths. Relative positional encodings, as used in *Transformer-XL* [15] and Rotary Position Embeddings (RoPE) proposed by Su et al. [16] provides better generalisation to longer contexts. More recent methods such as *YaRN* [17] adjust RoPE scaling to maintain performance across significantly extended context lengths.
- **Training and Fine-Tuning on Long Contexts**. Recent advancements show that increasing context length during pretraining can yield substantial improvements. Big models like Claude, Gemini and GPT-4 are examples of models trained or adapted for extended context windows up to 128k tokens or more. Techniques such as long-context fine-tuning, positional interpolation [18], and linear RoPE interpolation [7] have demonstrated effectiveness in scaling pretrained transformers to larger context windows without retraining from scratch.

The paper by Jin et al. [19] introduces a novel method called "*SelfExtend*", which enables LLMs to handle significantly longer contexts without any fine-tuning. This approach addresses the limitations of previous methods that often require extensive fine-tuning or architectural changes.

SelfExtend operates by constructing a bi-level attention mechanism during inference without modifying in any way the model structure or pretrained weights:

- **Neighbour Attention** focuses on dependencies among adjacent tokens within a specified range reducing the standard self-attention window to the closest positions. If L is the context window for the pretrained model, the parameter $w_n < L$

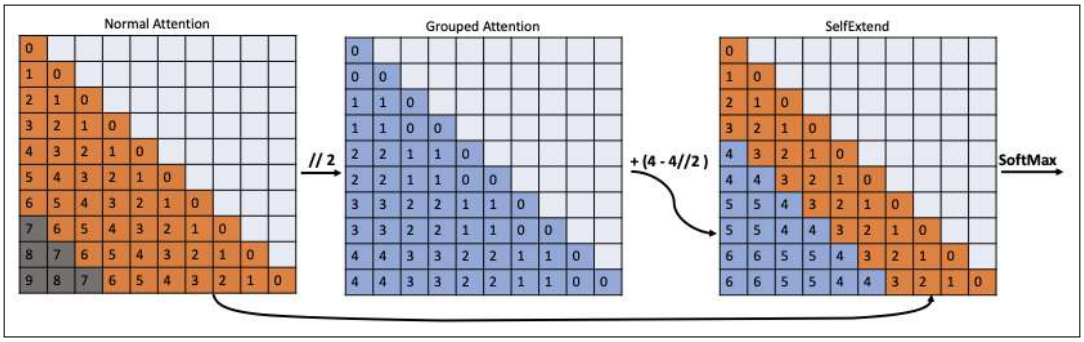

Figure 4: This figure shows the construction of the attention score matrix (before softmax) of SelfExtend: the example considers a sequence of length 10 fed into an LLM with the pretraining context window size $L = 7$. Numbers indicate the relative distances between the corresponding query and key tokens. Here $w_n = 4$ and $G_s = 2$. The two kind of attentions are then merged and the softmax operation is applied on the resulting matrix (picture and description taken from [19]).

controls the dimension of the neighbour attention.

- **Grouped Attention** captures dependencies among tokens that are far apart averaging the contributions of the pretrained self-attention between different G_s positions.

The maximum length of the extended context in the ideal case can be computed as

$$(L - w_n) * G_s + w_n \tag{1}$$

thus, for example, if we have $L = 4096$ and choose $w_n = 2048$ and $G_s = 16$, the ideal maximum extended context would be $34k$ tokens.

Figure 4 shows a small example of attention construction by mixing Neighbour and Grouped Attentions.

These two attention levels are computed based on the original model's self-attention mechanism, allowing for the extension of the context window with only minor code modifications and no need for additional training.

The authors argue that LLMs inherently possess the capability to handle long contexts, and the primary challenge lies in the out-of-distribution (O.O.D.) issues related to positional encoding. To mitigate this, SelfExtend maps unseen large relative positions to those observed during pretraining, effectively addressing the positional O.O.D. problem.

Empirical evaluations in Jin et al. [19] demonstrate that SelfExtend substantially improves the long-context understanding ability of LLMs and, in some cases, even outperforms fine-tuning-based methods on tasks such as language modeling, synthetic long-context tasks, and real-world long-context tasks.

This method has been successfully applied to various models, including LLaMA-2, Mistral, SOLAR, and Phi-2, showcasing its versatility and effectiveness in extending context windows without compromising performance.

More details on SelfExtend can be found in the original paper [19].

4.1. Using SelfExtend to increase LLMs context length

The baseline model for our experiments is the largest model produced by the SapienzaNLP team: *sapienzanlp/Minerva-7B-base-v1.0* is a Mistral-based model configured with a 4096-tokens fixed context and without sliding window pretrained from scratch on Italian and English [20]. Building on this baseline, we extended its context using *SelfExtend* with varying values of w_n and G_s, resulting in several variants referred to as "*LongMinerva*". These extended models were then evaluated on the proposed BABILong-ITA benchmark.

Figure 5 presents the results obtained by applying SelfExtend with seven different combinations of w_n and G_s. The method proves to be quite effective, enabling context extension for the original Minerva model maintaining similar performance for contexts $\leq$4k. Notably, the LongMinerva variants with $w_n = 512$ or 1024 and $G_s = 16$ achieved satisfactory performance improvements, given the original performance at 0k. Considering that SelfExtend operates without requiring any additional training or fine-tuning, these results seem particularly promising.

5. Discussion & Conclusion

This paper introduced a new benchmark for evaluating the effective context length of LLMs in Italian. Based on a similar resource originally developed for English, we translated and manually cleaned the data to construct a reliable and meaningful Italian benchmark.

Our evaluation of several prominent LLMs capable of processing Italian validated the quality of the proposed

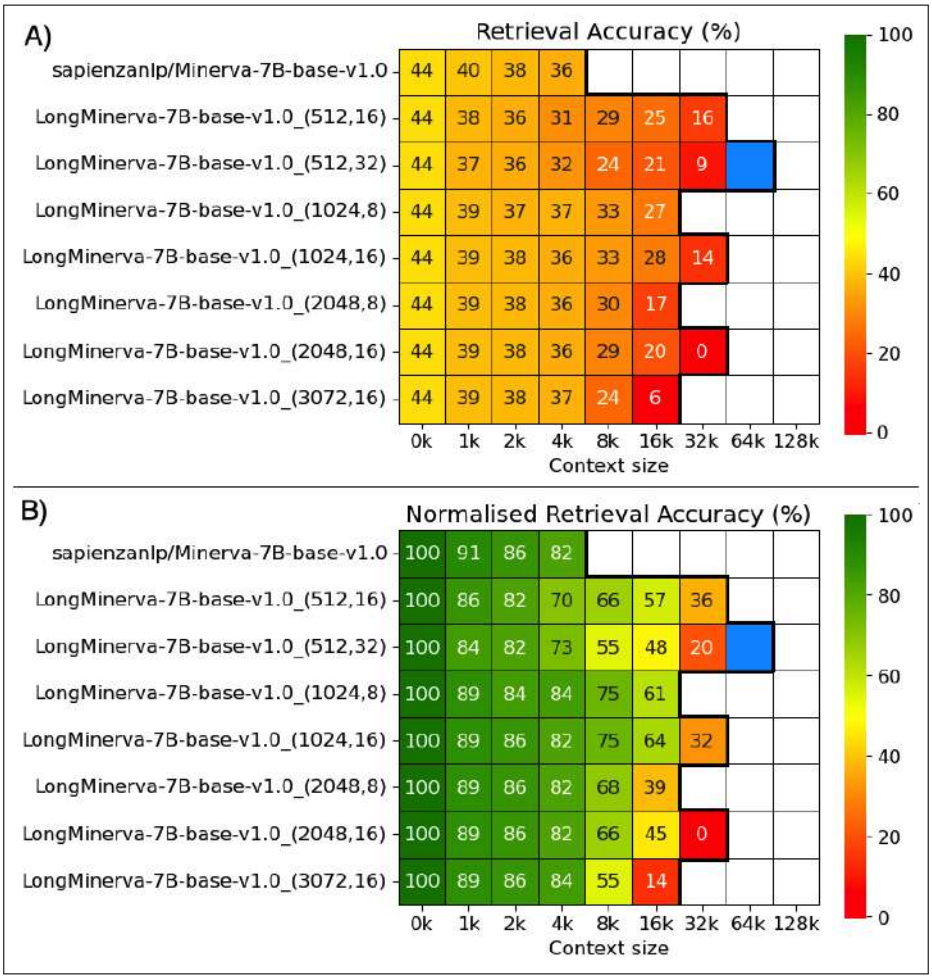

Figure 5: BABILong-ITA evaluation results for the experiments on context extension by using SelfExtend on the *sapienzanlp/Minerva-7B-base-v1.0* model, averaged over the five proposed tasks. In round brackets we have (w_n, G_s). The thick black line marks the maximum context length for a given extended context model computed using eq. (1). Blue squares indicate cells not computed for computational restriction reasons, but actually not useful for the evaluation because smaller contexts already presented very low retrieval accuracies. In A) we have original accuracies, while in B) we normalised the accuracy w.r.t. the "0k" case to show the relative reduction of accuracy obtained increasing the test context size.

benchmark and offered a clear picture of the actual context lengths these models can effectively handle.

The conclusions align closely with those reported in the original BABILong study by Kuratov et al. [9]: LLMs tend to struggle with retrieving relevant information at context lengths significantly shorter than their declared maximum capacities.

As an additional contribution, we applied the technique proposed by Jin et al. [19] to extend LLM context length without any training or fine-tuning, achieving promising results also for Italian large language models.

The benchmark data and all the codes for reproducing the experiments are available on Github[5].

Acknowledgments

I would like to thank the colleague S. Peroni for allowing me to use his GPU system to complete the experiments on extending LLM context length.

[5]https://github.com/ftamburin/BABILong-ITA

References

[1] T. Kočiský, J. Schwarz, P. Blunsom, C. Dyer, K. M. Hermann, G. Melis, E. Grefenstette, The NarrativeQA reading comprehension challenge, Transactions of the Association for Computational Linguistics 6 (2018) 317–328.

[2] Z. Yang, P. Qi, S. Zhang, Y. Bengio, W. Cohen, R. Salakhutdinov, C. D. Manning, HotpotQA: A dataset for diverse, explainable multi-hop question answering, in: E. Riloff, D. Chiang, J. Hockenmaier, J. Tsujii (Eds.), Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 2369–2380.

[3] W. Kryściński, N. Rajani, D. Agarwal, C. Xiong, D. Radev, Booksum: A collection of datasets for long-form narrative summarization (2021). arXiv:2105.08209.

[4] Y. Tay, M. Dehghani, S. Abnar, Y. Shen, D. Bahri, P. Pham, J. Rao, L. Yang, S. Ruder, D. Metzler, Long range arena : A benchmark for efficient transformers, in: International Conference on Learning Representations, 2021.

[5] Y. Bai, X. Lv, J. Zhang, H. Lyu, J. Tang, Z. Huang, Z. Du, X. Liu, A. Zeng, L. Hou, Y. Dong, J. Tang, J. Li, Longbench: A bilingual, multitask benchmark for long context understanding, 2024. arXiv:2308.14508.

[6] C. An, S. Gong, M. Zhong, X. Zhao, M. Li, J. Zhang, L. Kong, X. Qiu, L-eval: Instituting standardized evaluation for long context language models, in: L.-W. Ku, A. Martins, V. Srikumar (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 14388–14411.

[7] O. Press, N. Smith, M. Lewis, Train Short, Test Long: Attention with Linear Biases Enables Input Length Extrapolation, in: International Conference on Learning Representations, 2022.

[8] X. Wang, M. Salmani, P. Omidi, X. Ren, M. Rezagholizadeh, A. Eshaghi, Beyond the limits: a survey of techniques to extend the context length in large language models, in: Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI '24, 2024.

[9] Y. Kuratov, A. Bulatov, P. Anokhin, I. Rodkin, D. Sorokin, A. Sorokin, M. Burtsev, BABILong: Testing the Limits of LLMs with Long Context Reasoning-in-a-Haystack, in: A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, C. Zhang (Eds.), Advances in Neural Information Processing Systems, volume 37, Curran Associates, Inc., 2024, pp. 106519–106554.

[10] J. Weston, A. Bordes, S. Chopra, A. M. Rush, B. van Merriënboer, A. Joulin, T. Mikolov, Towards AI-Complete Question Answering: A Set of Prerequisite Toy Tasks, in: Proceedings of the International Conference on Learning Representations, 2016.

[11] J. W. Rae, A. Potapenko, S. M. Jayakumar, T. P. Lillicrap, Compressive transformers for long-range sequence modelling, in: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020, 2020.

[12] I. Beltagy, M. E. Peters, A. Cohan, Longformer: The long-document transformer, arXiv:2004.05150 (2020).

[13] M. Zaheer, G. Guruganesh, K. A. Dubey, J. Ainslie, C. Alberti, S. Ontanon, P. Pham, A. Ravula, Q. Wang, L. Yang, A. Ahmed, Big bird: Transformers for longer sequences, 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. 17283–17297.

[14] N. Kitaev, L. Kaiser, A. Levskaya, Reformer: The efficient transformer, in: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020, 2020.

[15] Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. Le, R. Salakhutdinov, Transformer-XL: Attentive language models beyond a fixed-length context, in: A. Korhonen, D. Traum, L. Màrquez (Eds.), Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 2019, pp. 2978–2988.

[16] J. Su, Y. Lu, S. Pan, A. Murtadha, B. Wen, Y. Liu, Roformer: Enhanced transformer with rotary position embedding (2021). arXiv:2104.09864.

[17] B. Peng, J. Quesnelle, H. Fan, E. Shippole, YaRN: Efficient context window extension of large language models, in: The Twelfth International Conference on Learning Representations, 2024.

[18] S. Chen, S. Wong, L. Chen, Y. Tian, Extending context window of large language models via positional interpolation (2023). arXiv:2306.15595.

[19] H. Jin, X. Han, J. Yang, Z. Jiang, Z. Liu, C.-Y. Chang, H. Chen, X. Hu, Llm maybe longlm: Selfextend llm context window without tuning, in: Proceedings of the 41st International Conference on Machine Learning, ICML'24, JMLR.org, 2024.

[20] 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.