EurekaRebus - Verbalized Rebus Solving with LLMs: A CALAMITA Challenge

Gabriele Sarti[1,*], Tommaso Caselli[1], Arianna Bisazza[1] and Malvina Nissim[1]

[1]Center for Language and Cognition (CLCG), University of Groningen, Oude Kijk in 't Jatstraat 26 Groningen, 9712EK, The Netherlands

Abstract

Language games can be valuable resources for testing the ability of large language models (LLMs) to conduct challenging multi-step, knowledge-intensive inferences while respecting predefined constraints. Our proposed challenge prompts LLMs to reason step-by-step to solve verbalized variants of rebus games recently introduced with the EurekaRebus dataset [1]. Verbalized rebuses replace visual cues with crossword definitions to create an encrypted first pass, making the problem entirely text-based. We introduce a simplified task variant with word length hints and adopt a comprehensive set of metrics to obtain a granular overview of models' performance in knowledge recall, constraints adherence, and re-segmentation abilities across reasoning steps.

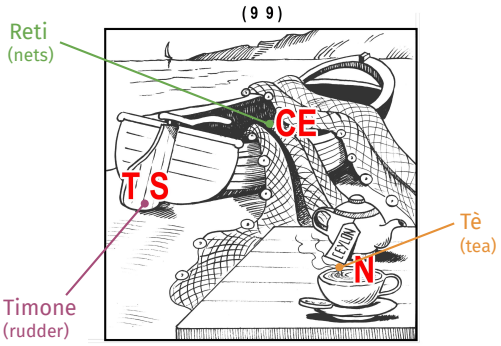
Keywords

Large language models, Sequential reasoning, Puzzle, Rebus, Crosswords, Enigmistica Italiana, CALAMITA

1. Challenge: Introduction and Motivation

Language games were adopted as testbeds for measuring NLP progress in recent years [2, 3, 4], with a particular focus on (cryptic) crossword solving English [5, 6, 7, 8, 9]. For the Italian language, initial efforts focused on crossword solving and generation [10, 11] and clue-based word guessing [12, 13, 9]. Recently, Sarti et al. [1] introduced an extensive collection of text-adapted Italian rebus puzzles to evaluate large language models' (LLMs) knowledge and sequential reasoning abilities. **Rebuses** are complex puzzles combining visual elements and graphic signs to encode a hidden phrase. Italian can boast a rich and long-standing rebus tradition dating back to the 19th century [14], popularized by high-diffusion magazines such as *La Settimana Enigmistica*[1]. The structure of Italian rebuses has, with time, been formalized into beauty canons [15], and their peculiarities and design principles were analyzed by several authors [16, 17, 18].

In Italian rebuses, rebus solving begins by combining derived by combining graphemes with their underlying visual elements in a left-to-right fashion, composing a **first pass** (*prima lettura*) representing an intermediate solution of the puzzle. Then, first pass elements are re-segmented (*cesura*) according to a **solution key** (*diagramma*), which specifies the length of each word in the **solution** (*frase risolutiva*). The **verbalized rebuses** introduced by Sarti et al. [1] are text-only version of real rebuses published in popular outlets derived by replacing words corresponding to visual elements with externally-sourced crossword definitions in the transcribed first passes, using a standardize format. Figure 1 provides a



First Pass: TeS timone - reti CE - N te

Verbalized Rebus:
TES [Dirige la rotta] *(Directs the course)*
[Le difendono i portieri] *(Protected by goalkeepers)* CE
N [Calda bevanda rilassante] *(Warm relaxing drink)*

Solution key (# of chars/word): 9 9

Solution: Testimone reticente *(reticent witness)*

Figure 1: Example of a verbalized rebus crafted by combining a rebus first pass (intermediate solution) with crossword definitions. Rebus by *Lionello*, art by Laura Neri.

CLiC-it 2024: Tenth Italian Conference on Computational Linguistics, Dec 04 — 06, 2024, Pisa, Italy

*Corresponding author.
g.sarti@rug.nl (G. Sarti); t.caselli@rug.nl (T. Caselli); a.bisazza@rug.nl (A. Bisazza); m.nissim@rug.nl (M. Nissim)
https://gsarti.com (G. Sarti); https://cs.rug.nl/~bisazza (A. Bisazza); https://malvinanissim.github.io (M. Nissim)
0000-0001-8715-2987 (G. Sarti); 0000-0003-2936-0256 (T. Caselli); 0000-0003-1270-3048 (A. Bisazza); 0000-0001-5289-0971 (M. Nissim)

[1]https://www.lasettimanaenigmistica.com/

simple example.

This work proposes to adopt the EurekaRebus introduced by Sarti et al. [1] to extend their evaluation of LLMs' multi-step reasoning and linguistic/cultural awareness to the systems evaluated as part of the CALAMITA evaluation campaign [19]. We believe the task is particularly relevant since the crossword definitions that compose verbalized rebuses rely heavily on idiomatic expressions, wordplay, and cultural references specific to Italian. Hence, the results of this task could provide valuable insights into the linguistic and cultural competence of LLMs trained on the Italian language. Moreover, the task is especially appealing since it is framed in a templated reasoning format, enabling us to disentangle the various components required to successfully solve a verbalized rebus step-by-step. More specifically, several metrics will be employed to assess LLMs' factual recall, textual concatenation and re-segmentation capabilities and, finally, constraint satisfaction given the provided cues.

In light of the results reported by [1] for state-of-the-art proprietary LLMs, we expect all tested open-source systems to perform very poorly, with final solution accuracies well below 30%. We also note that the highest reported overall performance in previous work[2] was found by the original authors to be primarily the product of memorization. We anticipate that this challenge will highlight significant limitations in LLMs' current factual recall and multi-step reasoning ability and act as a catalyst for future improvements in these areas.

2. Challenge: Description

The proposed challenge aims to evaluate the capabilities of existing LLMs in solving verbalized Italian rebuses via prompting at various granularity levels. More specifically, LLMs will be evaluated in a few-shot prompting setting with two fixed in-context learning examples pre-selected at random from the available pool of verbalized rebuses in EurekaRebus, in two settings:

- **Regular**, matching the example in table 1 and the original input format used by Sarti et al. [1].
- **Hints**, in which the number of characters for every hidden word is provided alongside definitions in the verbalized rebus to help the model in identifying the correct choice. This variant was not tested by Sarti et al. [1].

Refer to section 3.3 for the respective example formats. Models will be evaluated on their performance at each step required to successfully solve the verbalized rebus and their overall ability to produce correct final solutions.

[2]Namely 58% Solution Exact Match for a LLaMA-3.1 8B model LoRA-tuned on 80k EurekaRebus examples [20, 21]

3. Data description

3.1. Origin of data

The dataset used for this challenge is an extended version of **EurekaRebus** [1], a collection of 222,089 unique Italian rebuses extracted from Eureka5 platform[3], an open database of rebuses and other linguistic puzzles maintained by the Associazione Culturale "Biblioteca Enigmistica Italiana - G. Panini"[4]. Among these, 83,157 were converted by the original authors in verbalized form by leveraging the crossword definitions from the **ItaCW** collection [10], including 125,202 definition-solution pairs. While Sarti et al. [1] evaluated the performances of prompted and tuned LLMs on rebuses up to June 17th, 2024, the current test set include 168 new unseen examples released on Eureka5 after that date.

3.2. Annotation details

We employ the same procedure of Sarti et al. [1] for verbalizing available rebuses. More specifically, only rebuses having all lowercased or camel-cased words among ItaCW solutions are selected, and every word is replaced by sampling one of the available crossword definitions for it at random.[5] Moreover, only regular rebuses containing at least two hidden words are selected, avoiding examples requiring a single definition-solving step and those with more complex templates (e.g., *anarebuses* using anagrams of hidden words for the solution).

3.3. Data format

Each example in the dataset consists of:

- The **verbalized rebus** (`verbalized_rebus`) containing letters from the original rebus and crossword-style definitions enclosed in square brackets.
- A variant of the verbalized rebus containing **length hints** for definitions (`verbalized_rebus_with_length_hints`).
- The **solution key**, composed by whitespace-separated numbers representing the word lengths in the final solution (`solution_key`).
- The **first pass words** matching definitions in the verbalized rebus, provided in a semicolon-separated string in order of occurrence (`word_guesses`).
- The **first pass** obtained by infilling words in place of their definitions in the verbalized rebus (`first_pass`).

[3]http://www.eureka5.it
[4]http://www.enignet.it/home
[5]Words in ItaCW can be associated to multiple definitions.

```
{
    "verbalized_rebus": "[Edificio religioso] G [Lo fa doppio l'opportunista] NP [Poco cortese,
    ↪  severo] NZ [Parente... molto lontana]",
    "verbalized_rebus_with_length_hints": "[Edificio religioso (6)] G [Lo fa doppio l'opportunista
    ↪  (5)] NP [Poco cortese, severo (4)] NZ [Parente... molto lontana (3)]",
    "solution key": "3 1 6 3 8 2",
    "word_guesses": "chiesa;gioco;rude;ava",
    "first_pass": "chiesa G gioco NP rude NZ ava",
    "solution_words": "Chi;è;saggio;con;prudenza;va",
    "solution": "Chi è saggio con prudenza va"
}
```


Listing 1: Example entry for the challenge test set.

- The whitespace-separated **solution words** obtained after resegmenting the first pass according to the solution key, provided in a semicolon-separated string in order of occurrence (`solution_words`).
- The **solution** of the verbalized rebus used as the final prediction target for the LLM (`solution`).

An example is provided in Listing 1.

3.4. Prompting

Table 1 shows the 2-shot prompting template adopted for generating a templated solution with the tested LLMs. The second in-context example used in the template, omitted for brevity, corresponds to the one shown in Listing 1.

The task description provided to the model was derived from a trial-and-error process starting from the original prompt by Sarti et al. [1]. Notably, compared to the original authors the task description provides more detailed descriptions of individual components of the rebus to provide a clearer overview of the task to the LLM. We opted for a 2-shot setting as opposed to the 5-shot prompting employed by Sarti et al. [1] to accommodate the limited context length of some of the tested LLMs, thus ensuring that the total length after model generation does not exceed 1024 tokens[6]. The two examples provided remain the same shown here to simplify evaluation and ensure consistent results.

Verbalized rebus solving steps Table 1 provide labels for the steps necessary to solve the verbalized rebus that are considered in this challenge task. The model receives a **problem input** including a verbalized rebus (possibly with length hints) and a solution key (*chiave di lettura*). The first step involves resolving crossword definitions in order (**Definition resolution**), exploiting only the model's parametric knowledge to accomplish the task. Then, the resolved words need to be infilled into the original rebus to compose the first pass, and re-segmented in the **Solution segmentation** step. Finally, the individual solution words are reassembled into a single solution string.

[6]The LLaMA 3 tokenizer was used to perform this estimate

3.5. Detailed data statistics

Table 2 from Sarti et al. [1] reports statistics for the full and verbalized subsets of the EurekaRebus dataset.

Train set contents The training set contains 80,158 examples, which are ignored for the purpose of the CALAMITA campaign provided that no adaptation methods are evaluated.

Test set contents The test set contains 3,167 examples divided as follows, in order of appearance:

- 2000 examples matching the in-domain setting for models trained by [1], i.e. containing only first pass words seen by all available trained models.
- 999 examples matching the out-of-distribution setting for models trained by [1], i.e. containing at least one first pass word unseen during training by available trained models.
- 168 new verbalized rebuses added in EurekaRebus v1.1, added to the Eureka5 platform after June 17th, 2024. These can be either in-domain or out-of-distribution for models trained on the EurekaRebus's training set.

While prompted models should obtain similar performances across all test subsets, the aformentioned division will enable further comparisons with previously trained systems.

4. Metrics

The challenge employs a comprehensive set of metrics adapted from the original evaluation of [1]:

Prompt template	
Sei un'esperto risolutore di giochi enigmistici. Il seguente gioco contiene una frase (Rebus) nella quale alcune parole sono state sostituite da indizi tra parentesi quadre. I numeri in ogni indizio rappresentano la lunghezza della parola nascosta. Il tuo compito è quello di identificare le parole nascoste e sostituirle agli indizi nel Rebus, producendo una prima lettura dalla quale poi si deriverà una frase risolutiva. La chiave di lettura è una sequenza di numeri che rappresentano la rispettive lunghezze delle parole che compongono la frase risolutiva. La tua risposta deve essere una frase risolutiva sensata e che rispetti le lunghezze definite nella chiave di lettura.	
First example	# Esempio 1:
Problem input	*Rebus: AC [Un mollusco nell'insalata di mare (5)] GLI [Lo è l'operaio che lavora in cantiere (5)] S TO [Soldati da trincea (5)]* *Chiave di lettura: 11 2 10*
	Procediamo alla risoluzione del rebus passo per passo:
Definition resolution	*- A C = A C* *- [Un mollusco nell'insalata di mare] = cozza* *- G L I = G L I* *- [Lo è l'operaio che lavora in cantiere] = edile* *- S T O = S T O* *- [Soldati da trincea] = fanti*
First pass	*Prima lettura: AC cozza GLI edile S TO fanti*
	Ora componiamo la soluzione seguendo la chiave risolutiva:
Solution segmentation	*11 = Accozzaglie* *2 = di* *12 = lestofanti*
Solution	*Soluzione: Accozzaglie di lestofanti*
Second example	*# Esempio 2:* ... (same format as the first example)
Answer prefix	*# Ora tocca a te!* *Completa il rebus seguendo il procedimento descritto, rispondendo esattamente nello stesso formato utilizzato dagli esempi precedenti.* *Rebus:* ***{{verbalized_rebus}}*** *or* ***{{verbalized_rebus_with_length_hints}}*** *Chiave di lettura:* ***{{solution_key}}***

Table 1
2-shot prompt used for the CALAMITA evaluation. Blue text represent additions for the evaluation in the **Hints** setting. Template elements are highlighted next to the first in-context example. Example rebus by *Parodi E., Domenica Quiz n. 7*

Statistic	EurekaRebus	ItaCW-filtered
# examples	222089	83157
# authors	8138	5046
Year range	1800 - 2024	1869 - 2024
First pass		
# unique words	38977	8960
Avg./SD words/ex.	3.50/1/48	3.08/1.00
Avg./SD word len.	6.51/1.96	5.70/1.60
Avg./SD FP len.	26.45/11.19	25.74/8.73
Solution		
# unique words	75718	42558
Avg./SD words/ex.	3.02/1.60	2.80/1.21
Avg./SD word len.	8.07/2.30	7.79/2.23
Avg./SD Sol. len.	19.47/8.44	18.81/6.06

Table 2
Statistics for the full EurekaRebus dataset and the crosswords-filtered subset used in this work. Avg./SD = Average/standard deviation. Table adapted from Sarti et al. [1].

- **Word Guess Accuracy**: Proportion of correctly guessed words during definition resolution (corresponding to the Definition metric in the original evaluation).
- **Word Guess Length Accuracy**: Proportion of word guesses in definition resolution matching the correct length. This is evaluated only for the **Hints** setting, where the length is explicitly provided (not evaluated in previous works).
- **First Pass Accuracy**: Proportion of generated first passes matching the gold reference (corresponding to the First Pass Exact Match metric in the original evaluation).
- **Solution Word Accuracy**: Proportion of correct words in the generated solutions.
- **Solution Words Lengths Accuracy**: Proportion of generated solution words matching the lengths specified by the solution key. Lower scores may indicate difficulty in respecting the given length constraints (corresponding to the Solution Key Match metric in the original evaluation).
- **Solution Match**: Proportion of generated solutions matching the gold reference (corresponding to the Solution Exact Match metric in the original evaluation).

The Solution Match metric will be used as a primary metric of correctness, since it captures the model ability to fully solve the verbalized rebus. While no baseline evaluation was conducted for the new test set used in this challenge, we expect the performances of most capable open-source systems to align with those of 5-shot prompted LLaMA-3 70B and Qwen-2 72B models reported by Sarti et al. [1], which we summarize in Section 4. The results show that current models struggle

Model	Word Acc.	FP Acc.	Solution Word Acc.	Solution Word Len.	Solution Acc.
LLaMA-3 70B	0.22	0.04	0.03	0.16	0.00
Qwen-2 72B	0.28	0.04	0.04	0.20	0.00

Table 3
Baseline results for LLaMA-3 70B and Qwen-2 72B for the original test set, adapted from Sarti et al. [1].

to complete the task primarily due to incorrect word guesses, with errors propagating across resolution steps and ultimately resulting in a final accuracy of 0%.

5. Limitations

Several limitations should be considered when interpreting the results of this challenge:

Verbalization Simplification The use of verbalized rebuses, while necessary for text-based LLMs, simplifies the original visual puzzle. This does not fully capture the complexity of solving traditional rebuses, which rely on visual cues and cultural knowledge, making verbalized rebus solving a much simpler proxy to the multi-step reasoning required for regular rebuses.

Cultural Specificity The selected rebuses and crossword definitions rely heavily on Italian-specific linguistic and cultural background. Performance on this task may not generalize to other languages or puzzle types, and it might be unrealistic to expect general-purpose LLMs to possess the specific lexicon and knowledge used for rebus solving.

Prompt Sensitivity While the selected prompt template was observed to perform well for capable proprietary LLMs in preliminary tests, there are no guarantees that the instructions provided in the prompt are sufficient for smaller open-source models to perform verbalized rebus solving proficiently. Moreover, alternative prompt formulations could lead to potentially better results.

Lack of Human Baseline The challenge currently lacks a clear human performance baseline, which would be valuable for contextualizing model performance on verbalized rebus solving.

6. Ethical issues

While this challenge focuses on a relatively benign task of puzzle-solving, there are some ethical considerations to keep in mind. First, the dataset captures a very narrow subset of Italian language and culture. Hence, evaluation findings should not be overgeneralized to Italian language competence as a whole or to other cultures. This dataset's rebuses and crossword definitions are derived from commercially available published sources. While efforts have been made to ensure this data's exclusive, fair usage for research purposes, there may be copyright considerations to address.

7. Data license and copyright issues

As reported by the original EurekaRebus dataset license, the data is redistributed for research purposes only with the explicit approval of the Associazione Culturale "Biblioteca Enigmistica Italiana - G. Panini" (here onwards referred to as *the Association*), and the rights to each entry in the EurekaRebus collection are the property of the respective copyright holders. The usage and redistribution of these data is allowed only for users providing appropriate attribution to the original copyright holders and the Association, and the creation of derivative works is permitted only for research purposes, using terms no less restrictive than the EurekaRebus license. Researchers are encouraged to contact the challenge organizers with any questions or concerns about data usage and licensing.

Acknowledgments

We would like to express our gratitude to the following individuals and organizations:

- The Associazione Culturale "Biblioteca Enigmistica Italiana - G. Panini" for making their rebus collection freely accessible on Eureka5.
- The creators of the ItaCW dataset for enabling the creation of verbalized rebuses.
- The puzzle creators whose work is represented in this dataset.

Gabriele Sarti and Arianna Bisazza acknowledge the support of the Dutch Research Council (NWO) for the project InDeep (NWA.1292.19.399). Arianna Bisazza is further supported by the NWO Talent Programme (VI.Vidi.221C.009). We hope this challenge will contribute to the diffusion of the art of Italian *enigmistica* among computational linguistics and artificial intelligence researchers.

References

[1] G. Sarti, T. Caselli, M. Nissim, A. Bisazza, Non verbis, sed rebus: Large language models are weak solvers of italian rebuses, in: F. Dell'Orletta, A. Lenci, S. Montemagni, R. Sprugnoli (Eds.), Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024), CEUR.org, Pisa, Italy, 2024. URL: https://arxiv.org/abs/2408.00584.

[2] P. Giadikiaroglou, M. Lymperaiou, G. Filandrianos, G. Stamou, Puzzle solving using reasoning of large language models: A survey, ArXiv (2024). URL: https://arxiv.org/abs/2402.11291.

[3] B. J. Anderson, J. G. Meyer, Finding the optimal human strategy for wordle using maximum correct letter probabilities and reinforcement learning, Arxiv (2022). URL: https://arxiv.org/abs/2202.00557.

[4] G. Todd, T. Merino, S. Earle, J. Togelius, Missed connections: Lateral thinking puzzles for large language models, Arxiv (2024). URL: https://arxiv.org/abs/2404.11730.

[5] M. Ernandes, G. Angelini, M. Gori, Webcrow: A web-based system for crossword solving, in: AAAI Conference on Artificial Intelligence, 2005. URL: https://link.springer.com/chapter/10.1007/11590323_37.

[6] J. Rozner, C. Potts, K. Mahowald, Decrypting cryptic crosswords: Semantically complex wordplay puzzles as a target for nlp, in: M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, J. W. Vaughan (Eds.), Advances in Neural Information Processing Systems, volume 34, Curran Associates, Inc., 2021, pp. 11409–11421. URL: https://proceedings.neurips.cc/paper_files/paper/2021/file/5f1d3986fae10ed2994d14ecd89892d7-Paper.pdf.

[7] E. Wallace, N. Tomlin, A. Xu, K. Yang, E. Pathak, M. Ginsberg, D. Klein, Automated crossword solving, in: S. Muresan, P. Nakov, A. Villavicencio (Eds.), Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 3073–3085. URL: https://aclanthology.org/2022.acl-long.219. doi:10.18653/v1/2022.acl-long.219.

[8] A. Zugarini, K. Zeinalipour, S. S. Kadali, M. Maggini, M. Gori, L. Rigutini, Clue-instruct: Text-based clue generation for educational crossword puzzles, in: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), ELRA and ICCL, Torino, Italia, 2024, pp. 3347–3356. URL: https://aclanthology.org/2024.lrec-main.297.

[9] R. Manna, M. P. di Buono, J. Monti, Riddle me this: Evaluating large language models in solving word-based games, in: C. Madge, J. Chamberlain, K. Fort, U. Kruschwitz, S. Lukin (Eds.), Proceedings of the 10th Workshop on Games and Natural Language Processing @ LREC-COLING 2024, ELRA and ICCL, Torino, Italia, 2024, pp. 97–106. URL: https://aclanthology.org/2024.games-1.11.

[10] K. Zeinalipour, T. Iaquinta, A. Zanollo, G. Angelini, L. Rigutini, M. Maggini, M. Gori, Italian crossword generator: Enhancing education through interactive word puzzles, in: Proceedings of the 9th Italian Conference on Computational Linguistics (CLiC-it 2023), 2023. URL: https://ceur-ws.org/Vol-3596.

[11] G. Angelini, M. Ernandes, M. Gori, Solving italian crosswords using the web, in: International Conference of the Italian Association for Artificial Intelligence, 2005. URL: https://link.springer.com/chapter/10.1007/11558590_40.

[12] P. Basile, M. Lovetere, J. Monti, A. Pascucci, F. Sangati, L. Siciliani, Ghigliottin-ai@evalita2020: Evaluating artificial players for the language game "la ghigliottina" (short paper), EVALITA Evaluation of NLP and Speech Tools for Italian - December 17th, 2020 (2020). URL: https://doi.org/10.4000/books.aaccademia.7488.

[13] P. Basile, M. de Gemmis, P. Lops, G. Semeraro, Solving a complex language game by using knowledge-based word associations discovery, IEEE Transactions on Computational Intelligence and AI in Games 8 (2016) 13–26. doi:10.1109/TCIAIG.2014.2355859.

[14] D. Tolosani, Enimmistica, Hoepli, Milan, 1901.

[15] G. Brighenti, I canoni di bellezza nel rebus, Labirinto - Mensile di cultura enigmistica (1974). URL: http://win.cantodellasfinge.net/portale/leonardo/articoli/langense/pag2.asp.

[16] E. Miola, Che cos'è un rebus, Carocci, 2020.

[17] S. Bartezzaghi, Parole in gioco: Per una semiotica del gioco linguistico, Bompiani, 2017.

[18] P. Ichino, L'ora desiata vola: guida al mondo del rebus per solutori (ancora) poco abili, Bompiani, Milan, 2021.

[19] G. Attanasio, P. Basile, F. Borazio, D. Croce, M. Francis, J. Gili, E. Musacchio, M. Nissim, V. Patti, M. Rinaldi, D. Scalena, CALAMITA: Challenge the Abilities of LAnguage Models in ITAlian, in: Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024), Pisa, Italy, December 4 - December 6, 2024, CEUR Workshop Proceedings, CEUR-WS.org, 2024.

[20] M. AI, Introducing meta llama 3: The most capable openly available llm to date, Website, 2024. URL: https://ai.meta.com/blog/meta-llama-3.

[21] E. J. Hu, yelong shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, LoRA: Low-rank adaptation of large language models, in: The Tenth International Conference on Learning Representa-

tions (ICLR 2022), OpenReview, Online, 2022. URL:
https://openreview.net/forum?id=nZeVKeeFYf9.