Non Verbis, Sed Rebus: Large Language Models are Weak Solvers of Italian Rebuses

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

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

Abstract

Rebuses are puzzles requiring constrained multi-step reasoning to identify a hidden phrase from a set of images and letters. In this work, we introduce a large collection of verbalized rebuses for the Italian language and use it to assess the rebus-solving capabilities of state-of-the-art large language models. While general-purpose systems such as LLaMA-3 and GPT-4o perform poorly on this task, ad-hoc fine-tuning seems to improve models' performance. However, we find that performance gains from training are largely motivated by memorization. Our results suggest that rebus solving remains a challenging test bed to evaluate large language models' linguistic proficiency and sequential instruction-following skills.

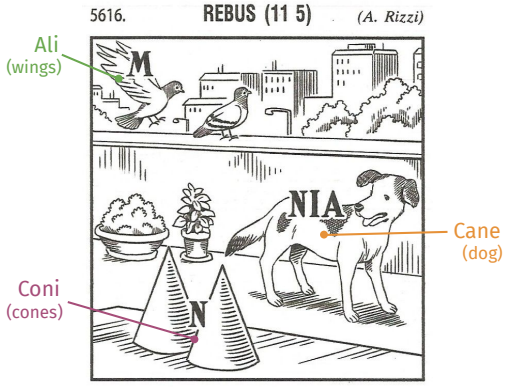
Keywords

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

1. Introduction

Complex games such as chess and Go have long been a source of inspiration to develop more flexible and robust AI systems [1, 2]. Recent developments in NLP suggested that creative language games could be exploited as promising benchmarks for quantifying the ability of large language models (LLMs) to carry out multi-step knowledge-intensive reasoning tasks under pre-specified constraints [3]. While crossword puzzles have been historically the main focus of such efforts [4], other categories of linguistic games received only marginal attention, especially for languages other than English. A prominent example of less-studied language games is the **rebus**, a visual puzzle combining images and graphic signs to encode a hidden phrase. Indeed, rebus solving is a complex, multi-step process requiring factual knowledge, contextual understanding, vocabulary usage, and reasoning within pre-defined constraints – a set of fundamental skills to address a variety of real-world tasks.

In this work, we conduct the first open evaluation of LLMs' rebus-solving capabilities, focusing specifically on the Italian language. We propose a novel strategy to derive text-only *verbalized rebuses* from transcribed intermediate rebus solutions and use it to produce a large collection with more than 80k verbalized rebuses. We then evaluate the rebus-solving skills of state-of-the-art LLMs, including open-source systems and proprietary models, via few-shot prompting. Moreover, we fine-tune a small but capable LLM on verbalized rebus solving, outperforming state-of-the-art systems by a wide margin. Finally, we conduct a fine-grained assessment of LLMs' sequential reasoning steps, explaining model performance in terms of word complexity and memorization.



First Pass: M ali - N coni - cane NIA

Verbalized Rebus:
M [Due calciatori attaccanti] *(Two attacking footballers)*
N [Usati per mangiare il gelato] *(Used for eating ice cream)*
[Abbaia e morde] *(Barks and bites)* NIA

Solution key (# of chars/word): 11 5

Solution: Malinconica nenia *(melancholic lullaby)*

Figure 1: An example of a verbalized rebus crafted by combining a rebus first pass (intermediate solution) with crossword definitions. We use verbalized rebuses to test LLMs' sequential instruction following capabilities. Image from *Settimana Enigmistica n. 4656*, © Bresi S.r.l.

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); m.nissim@rug.nl (M. Nissim); a.bisazza@rug.nl (A. Bisazza)
https://gsarti.com (G. Sarti); https://cs.rug.nl/~bisazza (A. Bisazza)
0000-0001-8715-2987 (G. Sarti); 0000-0003-2936-0256 (T. Caselli); 0000-0001-5289-0971 (M. Nissim); 0000-0003-1270-3048 (A. Bisazza)

Beyond rebus solving, our evaluation sheds light on the limits of current LLMs in multi-step reasoning settings, highlighting challenges with their application to complex sequential instruction-following scenarios.[1]

2. Background and Related Work

Italian *Enigmistica* and Rebuses The Italian language is characterized by a rich and long-standing tradition of puzzle games, including rebuses, dating back to the 19th century [5][2] In Italian rebuses, a **first pass** (*prima lettura*) representing an intermediate solution of the puzzle is produced by combining graphemes with underlying image elements in a left-to-right direction (Figure 1). Then, the letters and words of the first pass undergo a re-segmentation (*cesura*) according to a **solution key** (*chiave di lettura*[3]), which specifies the length of words in the **solution** (*frase risolutiva*). The **verbalized rebuses** we introduce in this work are variants of textual rebuses (*rebus descritto* or *verbis*), where the text-based puzzle is crafted by replacing first pass words with their crossword definitions in a templated format (Figure 1).

Linguistic Puzzles as NLP Progress Metrics Language games have recently been adopted as challenging tasks for LLM evaluation [3, 9, 10]. While works in this area have historically focused on English crosswords [11, 12, 4, 13], recent tests focus on a more diverse set of games such as the New York Times' "Connections" [14] and "Wordle" [15]. Automatic crossword solvers were also developed for French [16], German [17] and Italian [18, 19], while didactic crossword generators are available for Italian [20] and Turkish [21]. Relatedly, the Italian evaluation campaign EVALITA[4] recently hosted two shared tasks focusing on the word-guessing game "La Ghigliottina" (*The Guillotine*) [22, 23]. To our knowledge, our work is the first to attempt the computational modeling and evaluation of rebus-solving systems. Importantly, language games such as rebuses are not easily translatable into other languages due to their structural and cultural elements. This makes them a scarce but valuable resource for language-specific evaluations of language processing systems.

LLMs as Sequential Reasoners State-of-the-art LLMs were shown to struggle to follow sequential instructions presented in a single query [24], but their performances improved significantly with ad-hoc training [25]. This acts as an initial motivation for our rebus-solving fine-tuning experiments. In our evaluation, we also adopt few-shot prompting [26] and chain-of-thought reasoning [27], which were both shown to strongly improve LLMs' abilities when solving complex multi-step tasks.

3. Experimental Setup

Data We begin by extracting all rebuses' first passes and solutions available on Eureka5[5], an online repository of Italian puzzles. We refer to the resulting dataset containing 223k unique rebuses sourced from various publications as EurekaRebus. For crossword definitions, we use ItaCW [20], containing 125k unique definition-word pairs. We select only EurekaRebus examples in which all first pass words match an existing ItaCW definition to enable verbalization, maintaining 83,157 examples for our modeling experiments.[6] Since several ItaCW words are associated with multiple definitions, we randomly sample definitions to promote diversity in the resulting verbalized rebuses. A test set of 2k examples[7] is kept aside for evaluation, and the remaining 81k examples are used for model training.

Models We fine-tune Phi-3 Mini 3.8B 4K [28], the most capable LLM below 4B parameters for a wide range of Italian language tasks[8]. We use quantized low-rank adapters (QLoRA; 29, 30) for efficient fine-tuning with Unsloth[9] and Transformers [31], training the model for 5,000 steps with a batch size of 16 over 81k examples. For comparing our model performances, we select GPT-4o [32] and Claude-3.5 Sonnet [33] as the current state-of-the-art for proprietary LLMs and the instruction-tuned variants of Qwen-2 72B [34] and LLaMA-3 70B [35] as the best-performing open-source LLMs according to the Invalsi Italian benchmark [36]. These four systems are used as untrained baselines thanks to their instruction-following abilities and prompted for rebus solving in a few-shot setting.

Format Table 1 presents an example in the templated format used for fine-tuning Phi-3.[10] The model is prompted to reason step-by-step by 1) solving crossword definitions sequentially (**definition resolution**); 2) producing a **first pass** copying letters and definitions' words;

[1] Code, data and models are available on Github and Huggingface
[2] Refer to Miola [6], Bartezzaghi [7], Ichino [8] for a comprehensive overview of peculiarities and norms in modern Italian rebuses.
[3] Referred to as *diagramma* in jargon.
[4] https://www.evalita.it
[5] http://www.eureka5.it, additional details in Appendix A. Rebus illustrations are not available in Eureka5.
[6] Since verbalized rebus are produced from textual contents only, crossword definitions may refer to different word meanings (e.g. [Two soccer players] is used to represent the word "wings" in Figure 1 despite not matching the word sense "bird wings" of the original image. This does not affect the validity of our task.
[7] Composed by Test ID and Test OOD, described in Section 5
[8] https://hf.co/spaces/FinancialSupport/open_ita_llm_leaderboard
[9] https://github.com/unslothai/unsloth
[10] An English example is available in Table 9

Prompt	
Risolvi gli indizi tra parentesi per ottenere una prima lettura, e usa la chiave di lettura per ottenere la soluzione del rebus. *Rebus: U [Lo è il passacavallo] LO [È fatta di vimini] F F [Decimi di chilo] S [Disusato soprabito] A [Un rampicante dei Tropici]* *Chiave di lettura: 3 6 12 8*	
Generation	*Procediamo alla risoluzione del rebus passo per passo:*
Definition resolution	*- U = U* *- [Lo è il passacavallo] = nave* *- L O = L O* *- [È fatta di vimini] = cesta* *- F F = F F* *- [Decimi di chilo] = etti* *- S = S* *- [Disusato soprabito] = tait* *- A = A* *- [Un rampicante dei Tropici] = liana*
First pass	*Prima lettura: U nave LO cesta F F etti S tait A liana*
	Ora componiamo la soluzione seguendo la chiave risolutiva:
Solution segmentation	*3 = Una* *6 = veloce* *12 = staffettista* *8 = italiana*
Solution	*Soluzione: Una veloce staffettista italiana*

Table 1
Template used in fine-tuning and prompting experiments with highlighted reasoning stages. Example rebus by *Il Piacentino, Settimana Enigmistica n. 2942*

3) re-segmenting it into solution words based on the solution key (**solution segmentation**); and finally 4) producing the **solution** by copying re-segmented words. We automatically convert rebuses in this format by deriving the solution key from solution word lengths and dynamically infilling the available information into the template. We use a similar format for prompting experiments, with five in-context step-by-step demonstrations and an explicit instruction asking the model to stick to the previous examples' format to streamline solution parsing.

Metrics For our granular evaluation of rebus-solving performance, we adopt the following set of metrics focusing on the first passes (FP) and solutions (S) generated by LLMs:

- **Definition (Def.)**: Proportion of correctly guessed words during definition resolution.
- **First Pass Words/Letter Accuracy**: Proportion of correct words and letters in the generated first pass. Lower scores may indicate issues with assembling a first pass from previous information.
- **First Pass Exact Match (EM)**: Proportion of generated first passes matching the gold reference.
- **Solution Key Match**: Proportion of generated solution words matching the lengths specified by the solution key. Lower scores may indicate difficulty in respecting the given length constraints.
- **Solution First Pass Match**: Proportion of first pass characters employed to construct solution words. Lower scores indicate issues with using generated first pass characters in the solution.[11]
- **Solution Words Accuracy**: Proportion of correct words in the generated solution.
- **Solution Exact Match (EM)**: Proportion of generated solutions matching the gold reference.

4. Results

Table 2 presents our evaluation results. We observe that *all prompted models perform poorly on the task*, with the overall best prompted system (Claude 3.5 Sonnet) obtaining the correct solution only for 24% of the 2k tested examples. Notably, open-source systems perform significantly worse than proprietary ones, producing correct first passes only for 4% of the examples, and next to no correct solutions. Our fine-tuned system largely outperforms all state-of-the-art prompted models, predicting the correct solution in 51% of cases. From first pass metrics, it is evident these results can be largely explained by the poor word-guessing capabilities of the models, which are greatly improved with fine-tuning. For prompted models, the slight decrease in scores between Def. and FP Words also highlights issues with copying predicted words in the expected format. Finally, we observe that fine-tuning strongly improves the constraint-following abilities of our system, with prompted systems being less strict with applying length and letter-choice constraints for their solutions (Key/FP Match).

5. What Motivates Model Performances?

In light of the strong performances achieved by our relatively small fine-tuned system, this section conducts an in-depth investigation to identify factors motivating such performance improvements.

[11] In practice, we define this as $1 - \mathrm{CER}(\mathrm{FP}, \mathrm{S})$, where CER is the character error rate [37] between the two sequences (lowercased, whitespace removed) computed with Jiwer

Model	Setup	Def.	First Pass (FP)			Solution (S)			
			Words	Letters	EM	Key Match	FP Match	Words	EM
LLaMA-3 70B	5-shot prompt	0.22	0.20	0.60	0.04	0.16	0.51	0.03	0.00
Qwen-2 72B	5-shot prompt	0.28	0.25	0.76	0.04	0.20	0.52	0.04	0.00
GPT-4o	5-shot prompt	0.55	0.51	0.83	0.15	0.53	0.74	0.27	0.11
Claude-3.5 Sonnet	5-shot prompt	0.66	0.62	0.90	0.28	0.83	0.82	0.43	0.24
Phi-3 3.8B (ours)	fine-tuned	**0.84**	**0.84**	**1.00**	**0.56**	**0.86**	**0.94**	**0.68**	**0.51**

Table 2
Fine-grained verbalized rebus solving performances of various LLMs. **Bold** denotes best overall performances, and underline marks best training-free results.

Metric	GPT-4o			Phi-3 (ours)		
	Test ID	Test OOD	Test Δ	Test ID	Test OOD	Test Δ
FP W. ID	0.52	0.51	-0.01	0.96	0.96	0.00
FP W. OOD	-	0.44	-	-	0.20	-
FP EM	0.16	0.14	-0.02	0.89	0.18	-0.71
S W. ID	0.29	0.26	-0.03	0.92	0.49	-0.43
S W. OOD	0.18	0.16	-0.02	0.63	0.20	-0.40
S EM	0.12	0.09	-0.03	0.82	0.16	-0.66

Table 3
Model performances for test subsets containing only in-domain (Test ID), or some out-of-domain (Test OOD) first pass words. W. ID and W. OOD are accuracies for ID and OOD words for first pass (FP) and solution (S) sequences. Test Δ = Test ID - Test OOD performance.

Word Complexity and Frequency Affects LLM Fine-tuning Performance For every word in the first passes and solutions of test set examples, we measure LLMs' overall accuracy in predicting it for the full test set. We then correlate this score to various quantities that could motivate LLMs' performances. More specifically, we use 1) the word frequency in the training set; 2) the word frequency in PAISÀ [38], a large web Italian corpus; and 3) the length of the word (number of characters). We find a significant positive correlation ($\rho = 0.44$) between first pass word prediction accuracy and training frequency for the fine-tuned Phi-3 model, suggesting that model performance is strongly related to training coverage. The length of characters is also found to negatively affect our model's performance, albeit to a smaller extent ($\rho = -0.11$). The performance of prompted models is unrelated to both properties for first pass words, indicating that these results are the product of fine-tuning.[12]

LLM Fine-Tuning Fails to Generalize to Unseen Words To further confirm the importance of fine-tuning word coverage in defining model performances, we evaluate our fine-tuned model in out-of-distribution settings. For this evaluation, the 2k examples of the test set from previous sections are divided into two subsets: one in which all first pass words were seen during fine-tuning by Phi-3 (**Test ID**, 1061 examples) and one in which, for every example, at least one first pass word was unseen in training (**Test OOD**, 939 examples). Intuitively, if Phi-3 performance is mainly motivated by memorizing fine-tuning data, introducing OOD words should produce a significant drop in model performances. Results shown in Table 3 confirm that this is indeed the case. We find Phi-3 performances to be near-perfect on seen first pass words (FP W. ID = 0.96) in both test sets, with a major drop for OOD words (FP W. OOD = 0.20). This produces second-order effects on subsequent steps, causing the FP EM results to drop by 71% (FP EM Test Δ), while significantly impacting downstream solution accuracies. On the contrary, GPT-4o few-shot prompting performances remain nearly identical on both splits, confirming that these results are not the product of a skewed data selection process. Overall, these results strongly suggest that memorization is the main factor behind the strong rebus-solving performance of our fine-tuned LLM.

Manual Inspection We conclude by manually evaluating some generations produced by the best-performing LLMs. Table 4 presents two examples with definitions (D) and solution (S) words predicted by three LLMs, with more examples provided in Appendix C. We use NAW as short-hand for "Not A Word" to mark nonsensical terms.

In the first example, Phi-3 correctly predicts all first pass and solution words. On the contrary, other models make several mistakes in the first pass, leading to incorrect solutions. Both prompted models tend to ignore first pass words when these cannot be assembled to form sensical, length-fitting solution words. For example, for D1 GPT-4o predicts p (NAW), which would lead to the solution word "SAPpTE" (NAW), but the S8 = "Spettacolo" (*show*) is predicted instead by the model). In particular, GPT-4o appears to prioritize grammatically correct solutions at the cost of ignoring first pass words and solution key length constraints, while Claude 3.5S

[12] PAISÀ frequency is never found to correlate significantly. Full correlation results are available in Table 6.

Rebus: *SAP [La porta della breccia]* ^{D1} *TE [La pinza del granchio]* ^{D2} *SBA [Si legge su alcuni orologi]* ^{D3} *G [Le sue coccole sono aromatiche]* ^{D4} *V [Un gioco con dadi e pedine]* ^{D5} *D [Sono verdi in gioventù]* ^{D6}
Chiave di lettura: ***8 3 2 12 7 5***

Step	GPT-4o	Claude 3.5S	Phi-3
D1	p	one	pia
D2	chela	chela	chela
D3	ora	data	data
D4	ginepro	lio	ginepro
D5	ludo	oca	oca
D6	acerbi	anni	anni
S8	Spettacolo	Saponate	Sappiate
S3	che	che	che
S2	fa	la	la
S12	sognare	sbadataggine	sbadataggine
S7	ogni	vocando	provoca
S5	sera	danni	danni

Soluzione: *SAPpiaTE che la SBAdataGgine proVoca Danni*

Rebus: *STU [Si salva otturandolo]* ^{D1} *S [Ha foglie seghettate]* ^{D2} *AL [Lo è l'operaio che lavora in cantiere]* ^{D3} *G [Un uomo... non all' altezza]* ^{D4}
Chiave di lettura: ***11 7 2 7***

Step	GPT-4o	Claude 3.5S	Phi-3
D1	tappo	falla	dente
D2	acero	ortica	aro
D3	edile	edile	edile
D4	nano	nano	nano
S11	Stupaccerone	Stufallassor	Studentesaro
S7	salendo	ticale	aledile
S2	al	di	gi
S7	genano	Legnano	nanano

Soluzione: *STUdenteSsa liceALe di LeGnano*

Table 4
Examples of LLM generations for rebuses by *Slam, Nuova Enigmistica Tascabile n. 2802* (top) and *Grizzly, Domenica Quiz n. 2* (bottom). Correct guesses and errors and denoted for predicted first pass definitions ($D_{1,\dots,N}$) and solution words (S_i, with i being the i-th solution key value).

shows an improved ability to follow these constraints, as confirmed by Key/FP Match results of Table 2.

In the second example, the first pass word D2 = salice (*willow*) is OOD for Phi-3. Consequently, the model produces the incorrect prediction aro (NAW), and the error is propagated to all solution words, as previously observed in the Test OOD column of Table 3. Prompted models also underperform in this example, with errors on D1 and D2 propagating to most solution words. However, we note that D1 and D2 incorrect predictions for Claude 3.5S satisfy the provided definitions, suggesting that access to more explicit information about the given constraints could further boost LLMs' performance on this task.

6. Discussion and Conclusion

This work introduced a verbalized rebus-solving task and dataset for evaluating LLMs' sequential instruction following skills for the Italian language. We crafted a large collection of 83k verbalized rebuses by combining rebus transcriptions with crossword definitions and used it to evaluate the rebus-solving skills of state-of-the-art LLMs. Our experiments revealed the challenging nature of this task, with even the most capable prompted models achieving only 24% accuracy on solutions.

While fine-tuning a smaller LLM dramatically improved performance to 51% solution accuracy, our analysis uncovered that these gains were largely driven by memorization and do not generalize to out-of-distribution examples. These results suggest important limitations in the generalization capabilities of current systems for sequential instruction following tasks. Our manual analysis further shows that LLMs seldom account for length constraints when solving definitions, despite the fundamental role of these cues in restricting the pool of possible words. These results suggest that search-based approaches accounting for constraints more explicitly might improve puzzle structure adherence, as previously shown by Chen et al. [39]. Other augmentation techniques employing LLM reformulation skills can also be explored to mitigate overfitting.

Future work in this area should focus on expanding similar evaluations to a wider set of languages, input modalities, and puzzle categories, creating a comprehensive benchmark to test LLMs' puzzle-solving skills. Importantly, the task of solving visual rebuses and their more convoluted variants[13] remains far beyond the current capabilities of vision-language models. Hence, solving these puzzles automatically can be considered an important milestone in developing multimodal AI systems for constrained multi-step reasoning tasks. Our results confirm that the challenging nature of rebuses, even in their verbalized form, makes this task valuable for assessing future progress in LLMs' linguistic proficiency and sequential reasoning abilities. Finally, our rebus-solving LLM can facilitate future interpretability work investigating the mechanisms behind factual recall and multi-step reasoning in transformer models [40].

Limitations Our analysis was limited to a relatively small set of models, and a single prompt template obtained after minimal tuning. Further experiments are needed to verify that memorization patterns after fine-tuning remain relevant for other model sizes, prompt formats, and training regimes, particularly for full-weight training approaches.

[13] For example, rebuses requiring first pass anagrams (*anarebus*) or dynamic relations derived from multi-scene analysis (*stereorebus*)

Acknowledgments

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 are grateful to the Associazione Culturale "Biblioteca Enigmistica Italiana - G. Panini" for making its rebus collection freely accessible on the Eureka5 platform, and to Valeriya Zelenkova for her valuable comments on the first version of this work. We also thank the CLiC-it 2024 reviewers for their valuable feedback.

References

[1] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, D. Hassabis, Mastering the game of Go with deep neural networks and tree search, Nature 529 (2016) 484–489. doi:10.1038/nature16961.

[2] D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, T. Lillicrap, K. Simonyan, D. Hassabis, A general reinforcement learning algorithm that masters chess, shogi, and go through self-play, Science 362 (2018) 1140–1144. doi:10.1126/science.aar6404.

[3] 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.

[4] 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.

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

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

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

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

[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] 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.

[11] M. L. Littman, G. A. Keim, N. Shazeer, A probabilistic approach to solving crossword puzzles, Artificial Intelligence 134 (2002) 23–55. URL: https://www.sciencedirect.com/science/article/pii/S000437020100114X. doi:https://doi.org/10.1016/S0004-3702(01)00114-X.

[12] 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.

[13] A. Boda, Sadallah, D. Kotova, E. Kochmar, S. Yao, D. Yu, J. Zhao, I. Shafran, T. L. Griffiths, Y. Cao, K. N. 2023, S. Yousefi, L. Betthauser, H. Hasanbeig, R. Milliere, I. Momennejad, De-coding, A. Zugarini, T. Röthenbacher, K. Klede, M. Ernandes, B. M. Eskofier, D. Z. 2023, Are llms good cryptic crossword solvers?, ArXiv (2024). URL: https://arxiv.org/abs/2403.12094.

[14] 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.

[15] 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.

[16] G. Angelini, M. Ernandes, T. laquinta, C. Stehl'e, F. Simoes, K. Zeinalipour, A. Zugarini, M. Gori, The webcrow french crossword solver, in: Intelligent Technologies for Interactive Entertainment, 2023. URL: https://link.springer.com/chapter/10.1007/978-3-031-55722-4_14.

[17] A. Zugarini, T. Rothenbacher, K. Klede, M. Ernandes, B. M. Eskofier, D. Zanca, Die rätselrevolution: Automated german crossword solving, in: Proceedings of the 9th Italian Conference on Computational Linguistics (CLiC-it 2023), 2023. URL: https://ceur-ws.org/Vol-3596.

[18] 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.

[19] A. Zugarini, K. Zeinalipour, S. S. Kadali, M. Maggini, M. Gori, L. Rigutini, Clue-instruct: Text-based clue generation for educational crossword puzzles, 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. 3347–3356. URL: https://aclanthology.org/2024.lrec-main.297.

[20] 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.

[21] K. Zeinalipour, Y. G. Keptig, M. Maggini, L. Rigutini, M. Gori, A turkish educational crossword puzzle generator, ArXiv abs/2405.07035 (2024). URL: https://arxiv.org/abs/2405.07035v2.

[22] 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.

[23] 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.

[24] X. Chen, B. Liao, J. Qi, P. Eustratiadis, C. Monz, A. Bisazza, M. de Rijke, The sifo benchmark: Investigating the sequential instruction following ability of large language models, 2024. URL: https://arxiv.org/abs/2406.19999. arXiv:2406.19999.

[25] H. Hu, S. Yu, P. Chen, E. M. Ponti, Fine-tuning large language models with sequential instructions, Arxiv (2024). URL: https://arxiv.org/abs/2403.07794.

[26] 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.

[27] J. Wei, X. Wang, D. Schuurmans, M. Bosma, b. ichter, F. Xia, E. Chi, Q. V. Le, D. Zhou, Chain-of-thought prompting elicits reasoning in large language models, in: S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, A. Oh (Eds.), Advances in Neural Information Processing Systems, volume 35, Curran Associates, Inc., 2022, pp. 24824–24837. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf.

[28] M. Abdin, S. A. Jacobs, A. A. Awan, J. Aneja, A. Awadallah, H. Awadalla, N. Bach, A. Bahree, A. Bakhtiari, J. Bao, H. Behl, A. Benhaim, M. Bilenko, J. Bjorck, S. Bubeck, Q. C. et al., Phi-3 technical report: A highly capable language model locally on your phone, Arxiv (2024). URL: https://arxiv.org/abs/2404.14219.

[29] 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 Representations (ICLR 2022), OpenReview, Online, 2022. URL: https://openreview.net/forum?id=nZeVKeeFYf9.

[30] T. Dettmers, A. Pagnoni, A. Holtzman, L. Zettlemoyer, Qlora: Efficient finetuning of quantized llms, in: A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, S. Levine (Eds.), Advances in Neural Information Processing Systems, volume 36, Curran Associates, Inc., 2023, pp. 10088–10115. URL: https://proceedings.neurips.cc/paper_files/paper/2023/file/1feb87871436031bdc0f2beaa62a049b-Paper-Conference.pdf.

[31] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, A. Rush, Transformers: State-of-the-art natural language processing, in: Q. Liu, D. Schlangen (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Online, 2020, pp. 38–45. URL: https://aclanthology.org/2020.emnlp-demos.6. doi:10.18653/v1/2020.emnlp-demos.6.

[32] OpenAI, Hello gpt-4o, Website, 2024. URL: https://openai.com/index/hello-gpt-4o.

[33] Anthropic, Claude 3.5 sonnet, Website, 2024. URL: https://www.anthropic.com/news/claude-3-5-sonnet.

[34] A. Yang, B. Yang, B. Hui, B. Zheng, B. Yu, C. Zhou, C. Li, C. Li, D. Liu, F. Huang, G. Dong, H. Wei, H. Lin, J. Tang, J. Wang, J. Yang, J. Tu, J. Zhang, J. Ma, J. Xu, J. Zhou, J. Bai, J. He, J. Lin, K. Dang, K. Lu, K. Chen, K. Yang, M. Li, M. Xue, N. Ni, P. Zhang, P. Wang, R. Peng, R. Men, R. Gao, R. Lin, S. Wang, S. Bai, S. Tan, T. Zhu, T. Li, T. Liu, W. Ge, X. Deng, X. Zhou, X. Ren, X. Zhang, X. Wei, X. Ren, Y. Fan, Y. Yao, Y. Zhang, Y. Wan, Y. Chu, Y. Liu, Z. Cui, Z. Zhang, Z. Fan, Qwen2 technical report, 2024. URL: https://arxiv.org/abs/2407.10671.

[35] 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.

[36] F. Mercorio, M. Mezzanzanica, D. Potertì, A. Serino, A. Seveso, Disce aut deficere: Evaluating llms proficiency on the invalsi italian benchmark, 2024. URL: https://arxiv.org/abs/2406.17535.

[37] A. Morris, V. Maier, P. Green, From wer and ril to mer and wil: improved evaluation measures for connected speech recognition., 2004.

[38] V. Lyding, E. Stemle, C. Borghetti, M. Brunello, S. Castagnoli, F. Dell'Orletta, H. Dittmann, A. Lenci, V. Pirrelli, The PAISÀ corpus of Italian web texts, in: F. Bildhauer, R. Schäfer (Eds.), Proceedings of the 9th Web as Corpus Workshop (WaC-9), Association for Computational Linguistics, Gothenburg, Sweden, 2014, pp. 36–43. URL: https://aclanthology.org/W14-0406. doi:10.3115/v1/W14-0406.

[39] L. Chen, J. Liu, S. Jiang, C. Wang, J. Liang, Y. Xiao, S. Zhang, R. Song, Crossword puzzle resolution via monte carlo tree search, Proceedings of the International Conference on Automated Planning and Scheduling 32 (2022) 35–43. URL: https://ojs.aaai.org/index.php/ICAPS/article/view/19783. doi:10.1609/icaps.v32i1.19783.

[40] J. Ferrando, G. Sarti, A. Bisazza, M. R. Costa-jussà, A primer on the inner workings of transformer-based language models, Arxiv (2024). URL: https://arxiv.org/abs/2405.00208.

[41] C. Bonferroni, Teoria statistica delle classi e calcolo delle probabilita, Pubblicazioni del R. Istituto Superiore di Scienze Economiche e Commericiali di Firenze 8 (1936) 3–62.

A. Additional Data Information

Dataset statistics Table 5 presents statistics for the EurekaRebus dataset and the filtered subset we use for composing verbalized rebuses. The ItaCW dataset contains a total of 125,202 definitions for 40,963 unique words, with the most frequent words having hundreds of different definitions, e.g. 173 for *re* (king), 155 for *te* (you). Definitions used for verbalization are randomly sampled from the pool of available definitions for every word.

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 5
Statistics for the full EurekaRebus dataset and the crosswords-filtered subset used in this work. Avg./SD = Average/standard deviation.

Model	# Char.	Paisà Freq.	Train Freq.
GPT-4o	-0.01	0.01	0.02
Claude-3.5	-0.02	-0.02	0.00
Phi-3 (ours)	**-0.11**	-0.05	**0.44**
GPT-4o	**-0.18**	**0.14**	**0.19**
Claude-3.5	**-0.15**	**0.08**	**0.13**
Phi-3 (ours)	-0.02	**0.08**	**0.22**

Table 6
Spearman's correlation with average word accuracies for metrics computed on first pass (top) and solution (bottom) words. **Bold scores** are significant with Bonferroni-corrected $p < 1e-5$ [41]

First pass/Solution word distribution Figure 2 shows the distribution of first pass and solution words for the filtered EurekaRebus subset used in our work.

B. Additional Experimental Results

Table 6 presents the correlations between model accuracy and the properties presented in Section 5. Table 7 presents the full ID/OOD performances for all tested models, showing consistent results with Table 3 for all prompted models. Table 8 presents Phi-3 Mini performances across rebus-solving fine-tuning steps.

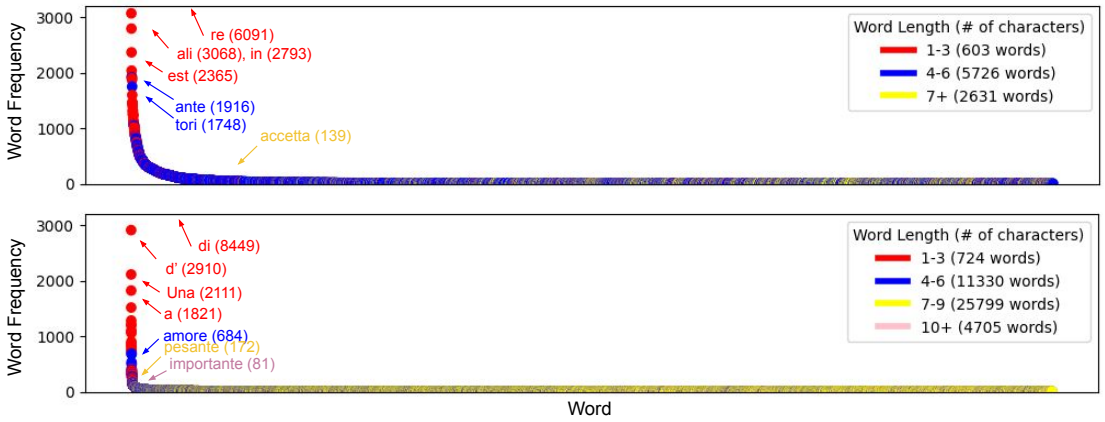


Figure 2: Word frequencies for words in first passes (top) and solutions (bottom) for the selected subset of EurekaRebus used for training and evaluation. Words are colored according to their length, and the most frequent examples per frequency bin are highlighted.

Metric	LLaMA-3			Qwen-2			GPT-4o			Claude-3.5S			Phi-3 (ours)		
	Test ID	Test OOD	Test Δ	Test ID	Test OOD	Test Δ	Test ID	Test OOD	Test Δ	Test ID	Test OOD	Test Δ	Test ID	Test OOD	Test Δ
FP W. ID	0.20	0.19	-0.01	0.26	0.25	-0.01	0.52	0.51	-0.01	0.65	0.63	-0.02	0.96	0.96	0.00
FP W. OOD	-	0.18	-	-	0.24	-	-	0.44	-	-	0.54	-	-	0.20	-
FP EM	0.03	0.04	0.01	0.03	0.05	0.02	0.16	0.14	-0.02	0.30	0.25	-0.05	0.89	0.18	-0.71
S W. ID	0.03	0.04	0.01	0.04	0.05	0.01	0.29	0.26	-0.03	0.48	0.40	-0.08	0.92	0.49	-0.43
S W. OOD	0.01	0.00	-0.01	0.02	0.00	-0.02	0.18	0.16	-0.02	0.41	0.30	-0.11	0.63	0.20	-0.40
S EM	0.00	0.00	0.00	0.00	0.00	0.00	0.12	0.09	-0.03	0.27	0.22	-0.05	0.82	0.16	-0.66

Table 7
Full model performances for test subsets containing only in-domain (Test ID), or some out-of-domain (Test OOD) first pass words. W. ID and W. OOD are accuracies for ID and OOD words for first pass (FP) and solution (S) sequences. Test Δ = Test ID - Test OOD performance.

C. Additional Model Generations

Table 9 presents an English translation of Figure 1 example using the prompt format adopted in this study.

Tables 10 and 11 provide additional example of LLM generations for tested rebuses, with the example from Table 11 (bottom) being OOD due to the *manovella* (crank) word in D2, and the others being ID for the fine-tuned

# Train Steps	Def.	First Pass (FP)			Solution (S)			
		Words	Letters	EM	Key Match	FP Match	Words	EM
500	0.64	0.63	0.97	0.25	0.66	0.86	0.36	0.16
1000	0.74	0.74	1.00	0.38	0.72	0.89	0.48	0.28
1500	0.78	0.77	0.99	0.42	0.78	0.91	0.55	0.34
2000	0.80	0.79	1.00	0.47	0.81	0.93	0.59	0.40
2500	0.81	0.81	1.00	0.49	0.81	0.92	0.62	0.42
3000	0.82	0.82	1.00	0.51	0.82	0.92	0.63	0.44
3500	0.82	0.82	1.00	0.53	0.83	0.93	0.65	0.47
4000	0.83	0.83	1.00	0.53	0.85	0.93	0.66	0.47
4500	0.83	0.83	1.00	0.55	0.85	0.93	0.67	0.49
5072 (final)	**0.84**	**0.84**	**1.00**	**0.56**	**0.86**	**0.94**	**0.68**	**0.51**

Table 8
Fine-grained verbalized rebus solving performances of the Phi-3 Mini model on the test set including both ID and OOD examples at various fine-tuned stages. **Bold** denotes best overall performances.

Prompt	
Solve the clues provided between parentheses to obtain a first pass, and use the solution key to obtain the rebus' solution. *Rebus:* M *[Two attacking footballers]* N *[Used for eating ice-cream] [Barks and bites]* NIA *Solution key: 11 5*	
Generation	*Let's proceed to solve the rebus step by step:*
Definition resolution	- *M = M* - *[Two attacking footballers]* = ali *(wings)* - *N = N* - *[Used for eating icecream]* = coni *(cones)* - *[Barks and bites]* = cane *(dog)* - *N I A = N I A*
First pass	*First pass: M ali N coni cane NIA*
	Let's now compose the solution by following the solution key:
Solution segmentation	*11* = Malinconica *(melancholic)* *5* = nenia *(lullaby)*
Solution	*Solution:* Malinconica nenia *(melancholic lullaby)*

Table 9
English translation of the template used in fine-tuning and prompting experiments with highlighted reasoning stages, with translated first pass and solution words in parentheses. Example rebus by *A. Rizzi, Settimana Enigmistica n. 4656*

Rebus: MOR [Musa della commedia] ^{D1} *L [Si accendono per vedere]* ^{D2} *NO [Uomini di intelligenza superiore]* ^{D3}

Chiave di lettura: ***7 12***

Step	**GPT-4o**	**Claude 3.5S**	**Phi-3**
D1	Talia	Talia	Talia
D2	luci	luci	luci
D3	geni	geni	genii
S**7**	Mortali	Mortali	Mortali
S**12**	allucinogeni	allucinogeni	allucinogeni

Soluzione: MORTali aLluciNOgeni

Table 10
Examples of LLM generations for a rebus by *De Vico C., Domenica Quiz n. 5.* Correct guesses and errors and denoted for predicted first pass definitions ($D_{1,\dots,N}$) and solution words (S_i, with i being the i-th solution key value).

Phi-3 Mini.

Rebus: PRI [Ricoperto di sudore] ^{D1} *MIN [Gli altari del tempio]* ^{D2} *DO [Un ordigno bellico]* ^{D3} *[Possono essere "di serie" in certi tornei]* ^{D4} *SSO*
Chiave di lettura: ***5 2 8 6 2 6***

Step	**GPT-4o**	**Claude 3.5S**	**Phi-3**
D1	sudato	madido	madido
D2	are	are	are
D3	bomba	mina	mina
D4	teste	teste	teste
S**5**	Prima	Prima	Prima
S**2**	di	di	di
S**8**	sudare	minaccia	dominare
S**6**	molto	teste	dominate
S**2**	di	di	se
S**6**	testa	dosso	stesso

Soluzione: PRIma di doMINare DOmina te steSSO

Rebus: AT [Si alzano nel camping] ^{D1} *[Emoziona pescatori e navigatori]* ^{D2} *[Come una nota Foresta]* ^{D3} *MEN [Quadro ad olio]* ^{D4} *S [Atteggiamento da modella]* ^{D5}
Chiave di lettura: ***9 11 2 5***

Step	**GPT-4o**	**Claude 3.5S**	**Phi-3**
D1	tende	tende	tende
D2	marea	mare	rete
D3	nera	nera	nera
D4	dipinto	tela	tela
D5	posa	posa	posa
S**9**	Attenderemo	Attendere	Attendere
S**11**	mareanera	marenamente	teneramente
S**2**	di	la	la
S**5**	posa	posa	sposa

Soluzione: ATtendere teneraMENte la Sposa

Rebus: B [Una figura geometrica] ^{D1} *[La si impugna per far girare un congegno]* ^{D2} *DA [Le produce il rovo]* ^{D3}

Chiave di lettura: ***10 7 1' 5***

Step	**GPT-4o**	**Claude 3.5S**	**Phi-3**
D1	cerchio	rombo	ellissi
D2	manovella	manovella	leva
D3	more	more	more
S**10**	Bcerchiomanovella	Bromomanov	Bellissile
S**7**		elladam	vadamore
S**1'**	d'	o'	'
S**5**	amore	more	remo

Soluzione: Bellissima novella D' Amore

Table 11
Examples of LLM generations for rebuses by *Baruffa, Rebus n. 12* (top), *Contini C., La Settimana Enigmistica n. 4102* (mid) and *Liosca, La Settimana Enigmistica n. 4581* (bottom). Correct guesses and errors and denoted for predicted first pass definitions ($D_{1,\dots,N}$) and solution words (S_i, with i being the i-th solution key value).