ProverbIT - Easy to complete, hard to choose: A CALAMITA Challenge

Enrico Mensa[1,*,†], Lorenzo Zane[2,†], Calogero J. Scozzaro[1], Matteo Delsanto[1], Tommaso Milani[2] and Daniele P. Radicioni[1]

[1]*Department of Computer Science, University of Turin, Turin, Italy*
[2]*Independent Researcher*

Abstract

We present ProverbIT, a benchmark designed to evaluate the reasoning capabilities of Large Language Models (LLMs) beyond simple pattern matching. While current models demonstrate high proficiency in text generation, their ability to discriminate between plausible but incorrect options remains understudied. ProverbIT addresses this gap through a challenging multiple-choice task focused on Italian proverbs. In this setting, models are provided with the beginning of a proverb and must select the correct completion from five options. Crucially, four options are always incorrect distractors, making the fifth option, 'None of the others', the only valid answer. This adversarial design forces models to abandon surface-level heuristics and engage in deeper semantic reasoning to actively discard misleading alternatives. To distinguish between a lack of knowledge and a failure in discriminative reasoning, we also introduce a generative completion baseline, where models simply complete the proverb from its initial fragment. The dataset comprises 100 common Italian proverbs, curated and validated by native speakers.

Keywords

Large Language Models, Reasoning, Multi-choice questions, Italian Proverbs

1. Challenge: Introduction and Motivation

The emergence of Large Language Models (LLMs) and Large Reasoning Models (LRMs) has revolutionized the natural language processing landscape across diverse domains [1]. Yet, while these models exhibit remarkable proficiency handling sophisticated linguistic phenomena [2], substantial uncertainty remains regarding their reliability in processing and interpreting culturally embedded linguistic expressions [3], such as proverbs. A proverb is a short, commonly known saying: it expresses a general truth, piece of wisdom, or practical advice, often based on common sense or cultural experience. The understanding of proverbs thus represents a key milestone in language proficiency, and access to the individual components of a proverb allows for the investigation of both lexical access issues and deeper semantic mechanisms.

Since proverbs are high-frequency patterns, standard completion tasks often yield high performance. While we assess this generative baseline, we moved beyond introducing a more complex challenge, evaluating discriminative selection. In this multiple-choice setting, the model must not only recognize the pattern but also evaluate and dismiss plausible alternatives.

In this work we introduce ProverbIT to the "Challenge the Abilities of LAnguage Models in ITAlian" (CALAMITA) initiative [4, 5]. ProverbIT is a dataset comprising multiple-choice questions centered on Italian proverbs presented at *CLiC-it 2025* [6, 7]. By manually designing alternative endings for the proverbs, we can systematically categorize error types and patterns. Our findings reveal a significant performance dichotomy: despite demonstrating some familiarity of these proverbs in generative settings, all models exhibit a sharp decline in accuracy when forced to operate within a multiple-choice framework.

EVALITA 2026: 9th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, Feb 26 – 27, Bari, IT
*Corresponding author.
†These authors contributed equally.
enrico.mensa@unito.it (E. Mensa); lorenzozane98@gmail.com (L. Zane); calogerojerik.scozzaro@unito.it (C. J. Scozzaro); matteo.delsanto@unito.it (M. Delsanto); milani.tommaso2004@gmail.com (T. Milani); daniele.radicioni@unito.it (D. P. Radicioni)
© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

2. Challenge Description

The challenge constitutes mainly a *multiple-choice* task, designed to assess the models' ability to select the correct proverb completion rather than relying on surface-level pattern matching. Specifically, given the beginning of a proverb, models are prompted to choose among several alternative endings that are all syntactically well-formed and, to varying degrees, semantically or stylistically plausible, but never correct. The task thus requires the models to actively evaluate and discard misleading options, identifying 'None of the others' as the only correct choice. This setting shifts the focus from automatic generation to discriminative reasoning, emphasizing fine-grained semantic control and proverb-level understanding.

To contextualize the results of the main task, we also introduce a generative *completion* task as a baseline. In this evaluation, models are asked to directly complete each proverb given its initial fragment. This baseline serves to estimate the models' prior familiarity with the proverbs and to verify whether potential errors in the multiple-choice setting stem from a lack of knowledge rather than from the nature of the selection task itself.

3. Data description

The ProverbIT dataset[1] is composed of 100 multi-choice questions, each regarding the completion of a specific Italian proverb. To create the dataset, we started from an initial set of 200 common Italian proverbs [8] from which we selected 100 of the most commonly used. This process was carried out by three of the authors, which are all native Italian speakers. Each proverb was then manually split into its *beginning* and its *ending*, with the point of division determined to maintain the proverb's semantic coherence in the initial part while allowing for a clear, unambiguous completion. For each proverb, four distinct incorrect alternative endings were manually created, leveraging the following constructive rationale:

- **A** is an ending that has similar sounds to the original continuation, often with an absurd/nonsensical meaning.
- **B** is a non assonant literal synonym of the original ending.
- **C** is the inverse of the original proverb ending, trying to maintain the assonance when possible.
- **D** is a tautological/trivial ending of the proverb, with no assonance.

For sake of clarity we provide an example in English for each of the aforementioned continuations. Completions for the proverb Actions speak... louder than words could be:

A) prouder than swords
B) at higher volume compared to speech
C) quieter than words
D) when they do

As this example shows, the synonym ending is not built on the figurative meaning of the proverb, but it is the literal synonym of the original ending (e.g., at higher volume compared to speech rather than beyond what words can say). This design was adopted to ensure that models cannot simply rely on surface-level syntactic patterns but must engage in deeper semantic and contextual reasoning to identify the absence of the correct completion.

3.1. Data format

The datasetis organized in a comma-separated values format. Each line contains one *complete proverb*, its *beginning* and *ending* splitted, and the four handcrafted incorrectalternatives.

[1]The full dataset can be downloaded at https://huggingface.co/datasets/emensa/proverbIT.

3.2. Prompts

Two different prompts were devised for the *completion* and *multi-choice* tasks. For the completion baseline, we adopted a simple prompt that requires the model to directly complete a proverb given its beginning:

Completion Prompt Template (translated)

Complete the proverb exactly:

[*Proverb beginning*]...

Reply with the ending only, do not add further comments.

On the multi-choice prompt we specify that the proverb must be completed *exactly*. Since all provided endings are incorrect, we expect models to always answer E) None of the other answers:

Multi-Choice Prompt Template (translated)

Complete the proverb exactly by choosing from the following options (which have no typing errors) indicating only the letter.

[*Proverb beginning*]...
A) ...[*Assonant ending]*
B) ...[*Synonym ending*]
C) ...[*Inverse ending*]
D) ...[*Trivial ending*]
E) None of the other answers

Do not add comments, the possible answers are only A, B, C, D, E.

For sake of clarity we provide an Italian example [with translation] from the actual dataset.

Example of proverb from the dataset

A buon intenditor,... [To a wise man]
A) ...foche canore [singing seals]
B) ...zero chiacchiere [zero chatter]
C) ...molte parole [many words]
D) ...è chiaro tutto [everything is clear]
E) Nessuna delle altre risposte [None of the other answers]

4. Metrics

For both tasks we employ an accuracy metric which is computed differently depending on the task.

Completion Task. To automatically calculate the accuracy on the *completion* task we compute the edit distance[2] between the provided completion and the correct ending of the proverb.

Multi-Choice Task. The accuracy of the *multi-choice* task was defined as the ratio of correctly chose options (which is always E) over the multiple choices.

[2]The implementation from https://docs.python.org/3/library/difflib.html was employed.

5. Results, Challenge-Specific Insights and Lessons Learned

The evaluation results are reported in Table 1. Each model was prompted once for each of the 100-samples in ProverbIT and with a zero shot setting. We adopted a chat template, no system prompt and a temperature of 0.

Model	Completion Results		Multi-Choice Results	
	Accuracy	Correct/Total	Accuracy	Correct/Total
Llama-3.1-70B-Instruct	0.67	67/100	0.14	14/100
Llama-3.1-8B-Instruct	0.20	20/100	0.03	3/100
LLaMAntino-3-ANITA-8B-Inst-DPO-ITA	0.13	13/100	0.05	5/100
Minerva-7B-instruct-v1.0	0.13	13/100	0.00	0/100

Table 1
Results on both the completion and multi-choice tasks.

Performance across models. The performances are overall very low, with Llama 70B being the best performing model (0.14 accuracy). For this bigger model, by comparing the results on the completion tasks vs the multi-choice task, we can observe a sharp drop in performances (from 0.67 to 0.14), indicating that even when the model recognizes and is able to complete a proverb, it is not competent enough to discard the four wrong completions in the multi-choice setting.

While the performance gap between tasks persists for smaller models, their competence is already limited in the completion baseline. A qualitative analysis of the outputs reveals a significant deficiency in instruction following: these models are often unable to adhere to simple constraint of 'only answer with a letter'. This verbosity is strictly penalized by the edit distance metric, leading to near-zero accuracy scores.

Error Analysis Table 2 gives a detailed report of the answers chosen by each model. The error analysis reveal distinct error patterns and biases among the models. Both Llama 70B and Minerva 7B display a strong tendency to select the 'Synonym' distractor (Option B), choosing it 56 and 62 times, respectively. Conversely, Llama 8B exhibits a bias toward the 'Assonant' distractor (Option A), selecting it in 43 instances. Additionally, the table highlights a significant instruction-following issue for Minerva 7B, which produced 33 'Invalid' responses, unlike the other models which strictly adhered to the valid letter format.

Model	A	B	C	D	E	Invalid
Llama-3.1-70B-Instruct	10	56	13	7	14	0
Llama-3.1-8B-Instruct	43	20	20	14	3	0
Minerva-7B-instruct-v1.0	5	62	0	0	0	33
LLaMAntino-3-ANITA-8B-Inst-DPO-ITA	31	21	13	30	5	0

Table 2
Absolute number of chosen answers in the ProverbIT task. A = Assonant, B = Synonym, C = Inverse D = Trivial E = None of the others. E is always the correct answer.

Discussion. The drop in performances between the completion and the multi-choice setting indicates that the higher completion accuracy likely stems from surface-level pattern matching rather than deep semantic comprehension. Since the models fail to filter out wrong completions when presented with distractors, it appears that simple language modeling is insufficient and that more powerful reasoning models are required to solve this task.

6. Limitations

While ProverbIT serves as a robust benchmark, we also report a few limitations. The dataset is relatively small, consisting of only 100 proverbs. While this size is sufficient for a challenge-oriented benchmark, it could limits the statistical power of the evaluation and may reduce the robustness of the conclusions drawn from the results. The construction of the alternatives in the multi-choice setting was performed manually. While this design choice allows for fine-grained control over the types of alternative endings, it also makes the dataset harder to scale and extend to a larger number of proverbs. Finally, we observe that especially when testing with smaller models, our metrics require precise instruction following, and so string-matching metrics like edit distance inflict severe penalties on models that are unable to suppress conversational filler in favor of the requested output.

7. Ethical issues

Our challenge poses no ethical issues and the ProverbIT dataset does not contain sensitive topics that could lead to ethical issues such as regional differences or social status.

8. Data license and copyright issues

ProverbIT dataset is distributed under the Creative Commons License Attribution 4.0 International (CC BY 4.0).

Declaration on Generative AI

During the preparation of this work, the authors used gemini-3-pro-preview in order to: Grammar and spelling check. After using these tool, the authors reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

[1] A. Lewkowycz, A. Andreassen, D. Dohan, E. Dyer, H. Michalewski, V. Ramasesh, A. Slone, C. Anil, I. Schlag, T. Gutman-Solo, et al., Solving quantitative reasoning problems with language models, Advances in Neural Information Processing Systems 35 (2022) 3843–3857.

[2] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang, et al., A survey on evaluation of large language models, ACM transactions on intelligent systems and technology 15 (2024) 1–45.

[3] F. D. L. Fornaciari, B. Altuna, I. Gonzalez-Dios, M. Melero, A hard nut to crack: Idiom detection with conversational large language models, in: Proceedings of the 4th Workshop on Figurative Language Processing (FigLang 2024), 2024, pp. 35–44.

[4] G. Attanasio, P. Basile, F. Borazio, D. Croce, M. Francis, J. Gili, E. Musacchio, M. Nissim, V. Patti, M. Rinaldi, et al., Calamita: Challenge the abilities of language models in italian, in: Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024), Pisa, Italy, 2024.

[5] M. Nissim, D. Croce, V. Patti, P. Basile, G. Attanasio, E. Musacchio, M. Rinaldi, F. Borazio, M. Francis, J. Gili, D. Scalena, B. Altuna, E. Azurmendi, V. Basile, L. Bentivogli, A. Bisazza, M. Bolognesi, D. Brunato, T. Caselli, S. Casola, M. Cassese, M. Cettolo, C. Collacciani, L. D. Cosmo, M. P. D. Buono, A. Esuli, J. Etxaniz, C. Ferrando, A. Fidelangeli, S. Frenda, A. Fusco, M. Gaido, A. Galassi, F. Galli, L. Giordano, M. Goffetti, I. Gonzalez-Dios, L. Gregori, G. Grundler, S. Iannaccone, C. Jiang, M. L. Quatra, F. Lagioia, S. M. Lo, M. Madeddu, B. Magnini, R. Manna, F. Mercorio, P. Merlo, A. Muti, V. Nastase, M. Negri, D. Onorati, E. Palmieri, S. Papi, L. Passaro, G. Pensa, A. Piergentili, D. Potertì, G. Puccetti, F. Ranaldi, L. Ranaldi, A. A. Ravelli, M. Rosola, E. S. Ruzzetti, G. Samo, A. Santilli, P. Santin, G. Sarti, G. Sartor, B. Savoldi, A. Serino, A. Seveso, L. Siciliani, P. Torroni, R. Varvara,

A. Zaninello, A. Zanollo, F. M. Zanzotto, K. Zeinalipour, A. Zugarini, Challenging the abilities of large language models in italian: a community initiative, 2025. URL: https://arxiv.org/abs/2512.04759. arXiv:2512.04759.

[6] E. Mensa, L. Zane, C. J. Scozzaro, M. Delsanto, T. Milani, D. P. Radicioni, Easy to complete, hard to choose: Investigating LLM performance on the ProverbIT benchmark, in: C. Bosco, E. Jezek, M. Polignano, M. Sanguinetti (Eds.), Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025), CEUR Workshop Proceedings, Cagliari, Italy, 2025, pp. 722–734. URL: https://aclanthology.org/2025.clicit-1.69/.

[7] C. Bosco, E. Jezek, M. Polignano, M. Sanguinetti (Eds.), Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025), CEUR Workshop Proceedings, Cagliari, Italy, 2025. URL: https://aclanthology.org/2025.clicit-1.0/.

[8] F. Caramagna, I 200 proverbi italiani più belli e famosi (con significato), 2025. URL: https://aforisticamente.com/i-200-proverbi-italiani-piu-belli-e-famosi-con-significato/.