Subcategorization of Italian Verbs with LLMs and T-PAS

Luca Simonetti[1,2,*], Elisabetta Ježek[3] and Guido Vetere[1]

[1]Università degli studi Guglielmo Marconi, Via Plinio, 44, 00193 Roma, Italy
[2]Università degli studi di Udine, Via Palladio, 8, 33100 Udine, Italy
[3]Università di Pavia, Corso Strada Nuova, 65, 27100 Pavia, Italy

Abstract

This study explores the application of Large Language Models (LLMs) to verb subcategorization in Italian, focusing on the identification and classification of syntactic patterns in sentences. While LLMs have made lexical analysis more implicit, explicit argument structure identification remains crucial in domain-specific contexts. The research leverages T-PAS, a rich lexical resource for Italian verbs, to fine-tune the open multilingual model Mistral 7B using the Iterative Reasoning Preference Optimization (IRPO) technique. This approach aims to enhance the recognition and extraction of verbal patterns from Italian sentences, addressing challenges in resource quality, coverage, and frame extraction methods. By combining curated lexical-semantic resources with neural language models, this work contributes to improving verb subcategorization tasks, particularly for the Italian language, and demonstrates the potential of LLMs in refining linguistic analysis tools.

Keywords

NLP, T-PAS, Verb Subcategorization, Mistral, CLiC-it

1. Introduction

Verb subcategorization is the task of identifying and classifying the syntactic patterns (or frames) taken by verbs in sentences. These patterns encode the possible combinations of arguments (such as subjects, objects, and complements) that a verb can have, specifying the number and type of arguments as well as their syntactic and semantic roles. Verb subcategorization is often used in Natural Language Understanding (NLU) to provide the main interpretation backbone. Although recent developments brought about by Large Language Models (LLM) make lexical analysis somewhat implicit, there are cases in which the identification of the argument structure of the verb is required, especially those where extensive domain-specific knowledge is required.

Semantic lexical resources such as VerbNet[1], FrameNet[2] and PropBank[3] have been largely employed for several NLP tasks in the past decades, including accomplishing verbal framing for the English language. VerbNet, for example, has been used to improve semantic role labeling, verb sense disambiguation and ontology mapping ([4], [5]); its new enhanced semantic representations have also recently been used for entity state tracking [6]. The main problems addressed in these experiences concern the quality and coverage of such resources and the methods used to extract frames from sentences.

Neural Language Models can help address both these issues. On the one hand, they may facilitate the construction of curated lexical-semantic resources; on the other hand, they can power robust frame-sentence matching procedures. The present work focuses on the Italian language. It concerns an experiment of using a rich lexical resource for Italian verbs, namely T-PAS [7] to fine-tune an open multilingual model, namely `Mistral 7B` [8], to recognize and extract verbal patterns from Italian sentences using a technique called IRPO [9].

The paper is organized as follows: in Section 2 we introduce the T-PAS resource for Italian verbs, which we used in our experiments. Section 3 discusses in detail the methodology we applied and references closely related works, whereas Section 4 illustrates the experimental setup. We complete the paper by discussing our results in Section 5 and by drawing some conclusions as well as making suggestions for future research in Section 6.

2. The T-PAS resource

T-PAS [7] is an inventory of argument structures and senses for Italian verbs.[1] In T-PAS, for each verb meaning, a specific Typed Predicate-Argument Structure (T-PAS, informally called pattern) is provided, in which arguments are defined in terms of semantic classes notated between square brackets, called semantic types. An example of a pattern for the verb *guidare* 'drive' in its 'operate' sense is [Human] guida [Vehicle]. Patterns are acquired from corpora following the Corpus Pattern

[1]The T-PAS project was developed at the Department of Humanities of the University of Pavia, with the technical support of Lexical Computing Ltd. The resource can be freely accessed and downloaded at https://tpas.unipv.it.

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

*Corresponding author.
simonetti.luca@spes.uniud.it (L. Simonetti); jezek@unipv.it (E. Ježek); g.vetere@unimarconi.it (G. Vetere)
0000-0003-2518-5200 (E. Ježek); 0000-0002-6703-7276 (G. Vetere)

Analysis (CPA) methodology [10]. Currently, T-PAS contains 1160 analyzed verbs, 5529 patterns and ca. 200,000 annotated corpus instances. Semantic types (Human, Event, Location, Food, Vehicle, etc.) are obtained from manual clustering of the lexical items found in the argument positions in the corpus. These types look very much like ontological categories; however, instead of being stipulated, they are induced from corpus data and reflect how humans talk about events and states of entities through language. The system of semantic types in T-PAS currently contains 180 semantic types. The list is organized in a hierarchy to identify the appropriate level of specificity of the selectional properties of individual verbs.

3. Background and Methodology

The extraction of verbal frames consists of applying frame-like structures to sentences. Once a suitable frame is identified, each element of the structure is mapped to an element of the sentence. To start our experiment, we attempted to extract the frame directly from the neural model (LLM), relying on the fact that LLMs are pretrained on large amount of texts and that their language modeling capabilities have reached unprecedented levels of maturity in the last three years. Although promising, this approach proved insufficient, since the model struggled with the correct subcategorization of the verb before extracting the appropriate frame. In a way, it appeared that the selection was compromised by the nondeterministic nature of LLM inference. Consequently, we split the task into two separate phases: 1) frame identification, i.e. T-PAS subcategorization, and 2) frame extraction, i.e. frame-sentence mapping.

We found that the baseline model performed poorly on the subcategorization task, achieving only 59.8% accuracy. For this reason, we decided to fine-tune the baseline model on the task of identifying verbal frames, which proved to be key for the subsequent task of extracting these frames. This approach was inspired by [11], where the authors set up a framework for verb sense disambiguation by providing the model with the frame that describes the sense the verb can take. This allows us to treat this task as a linguistic and semantic task rather than a simple categorization task. The idea is to provide the model with a prompt that includes the frames, based on the hypothesis that supplying the model with as much information as possible might be beneficial. This paper will only cover the subcategorization task. To do this, we created a fine-tune dataset based on the T-PAS resource, containing both the necessary information and a large number of examples to build upon.

4. Experimental Setup

The experimental setup consists of two main stages: dataset creation and fine-tuning of the base model *Mistral 7B* [8], as per the paper Iterative Reasoning with Preference Optimization (*IRPO*) [9]. Our implementation involves a single iteration, comprising both dataset generation and the actual fine-tuning. Additionally, we conduct a basic fine-tuning process where we train Mistral to directly complete prompts with the correct answer in a specified format:

```
La risposta corretta è 2...
```

We refer to this as the SFT (Supervised Fine-Tuning) model later in the discussion. This approach allows us to compare the effectiveness of IRPO against a more straightforward fine-tuning method. We now provide more details about the two stages of our experimental setup.

4.1. Dataset Creation

The first stage, dataset creation, involves the following steps:

1. We collect 30 responses from the base Mistral Model with a *high temperature* for each sentence.
2. Using these responses, we build a dataset containing $(x_i, y_{w,i}, y_{l,i})$ tuples, where:
    - x_i is the prompt used in step 1 to generate the responses
    - $y_{w,i}$ is the winning response (i.e., the one that leads to a correct answer)
    - $y_{l,i}$ is the losing response (i.e., the wrong one)

The first phase involves gathering sentences and structuring prompts. The prompts consist of questions to the model, where we ask which of the listed senses is the correct one for the sentence we provide. We use a subset of the T-PAS dataset, comprising approximately 5,324 examples out of the total 26,652 elements (around 19.9% of the full dataset). The sentences are randomly picked from this subset, using at most two examples for each verb to avoid any bias towards one specific pattern or predicate. This approach ensures a diverse representation while maintaining a manageable dataset size for our experiments. The possible senses the verb can acquire are constructed from the T-PAS dataset. We maintain the original order of the senses as listed in T-PAS to facilitate both the dataset generation and the evaluation processes. Our preliminary tests indicated that this decision doesn't significantly affect performance. We provide an example of a prompt in the Appendix to illustrate the structure and content of our queries to the model. After building the prompts, we query the Mistral 7B model API 30 times with temperature set to the highest value to let the model explore as much as possible its internal latent space to provide a response. This second phase results in 30 responses per prompt. We then compile a dataset of both correct and incorrect responses. The prompt instructs the model to answer in a specific format. Even if a response is semantically correct but doesn't adhere to the required format, we classify it as incorrect and include it with the wrong responses. This approach aligns with the methodology of the base paper and serves multiple purposes in our tuning process. By enforcing a specific format, we're not only training the model to provide correct answers but also to follow instructions precisely which provides us with a standardized format that ensures consistency across responses, crucial for large-scale evaluation and comparison. This phase results in an unbalanced dataset of wrong and right responses for each prompt summing up to 30.

The second step of this pipeline involves transforming the intermediate dataset into the final dataset. Following the approach of the IRPO authors, we combine chosen and rejected responses to create a balanced dataset, ensuring that each response is processed at least once during fine-tuning. For example, if the number of elements in one of the two stacks – *chosen* and *rejected* – is less than the other, we reuse elements from the stack with fewer items multiple times to achieve balance.

The result is a dataset consisting of 17,863 rows with columns *prompt*, *chosen*, *rejected*, which we make available on Hugginface[2].

4.2. Fine-Tune

The loss function we employ in the second stage, the fine-tuning, consists of two components: one handling the Direct Preference Optimization (*DPO*) Rewards [12], and another that positively affects the Negative Log Likelihood (*NLL*) of the correct answer. This approach has similar effects to those described by the authors of IRPO [9].

Using the dataset built as described in the previous section we proceed with the fine-tune. We also build the custom loss function as described (but not implemented) by the authors of IRPO. Our implementation of the replicated loss function will be made publicly available. The LoRA [13] configuration is as follows:

```
rank=16,
lora_alpha=16,
lora_dropout=0.05,
bias="none",
task_type="CAUSAL_LM",
target_modules=[
'k_proj', 'gate_proj', 'v_proj',
'up_proj', 'q_proj', 'o_proj',
'down_proj']
```

[2]https://huggingface.co/datasets/theGhoul21/irpo-dataset-v2

We use a single L4 GPU chip with 24GB VRAM available. As such, we can only have a batch size of 1 and use gradient accumulation of 2 to simulate a batch size of 2. We set max steps to 4,500 steps but actually stop the fine-tuning after 3,000 steps since there seems to be a plateau in the performances of the reward accuracy (see Figure 1).

4.3. Differences from the IRPO paper

The main differences from the original IRPO setup are as follows:

1. As starting model M_0 we use Mistral 7B: in other words a different model with 10x less parameters whereas the authors of the original paper use Llama-2 70B, a different model with different architecture, and possibly different dataset used in the pretraining.
2. We apply IRPO to a linguistic task instead of logic or math reasoning task.
3. We use a subset of verbs for training and observe generalization on different verbs during testing. This approach differs from the IRPO authors, who utilize standardized datasets such as GSM8K, MATH, and ARC-Challenge. While these datasets allow for direct comparison across different models and techniques, they don't provide the opportunity to assess generalization to unseen problem types in the same way our verb-based approach does.

5. Results and Discussion

Our final results are summarized in Table 1. We observe that the multilingual baseline model, although including Italian, is not sufficiently accurate in performing the selection task. Furthermore, when the model's temperature is increased, it does not remain consistent with a specific answer but rather explores multiple response options, selecting a different choice randomly each time. This could be explained in multiple ways: one is that the model knows it needs to select an answer but doesn't relate to the correct one using a thorough analysis but rather following a *pick-one* strategy with the explanation coming as a consequence. The other possible explanation is that the model just tries to give an answer, not actually connecting pieces of the given possibilities to the sentence but rather picking random parts of the sentence where they are more likely to reside for that particular part of the frame – e.g. the subject is usually heading the sentence. Unfortunately these are only speculations

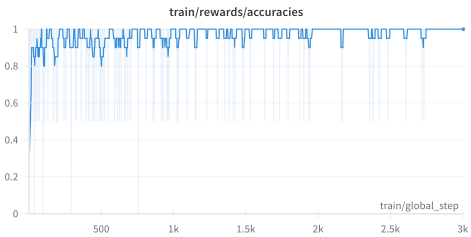

Figure 1: From Random to Expert: Rewards Accuracy of the Model Over Time. This graph illustrates the rapid learning curve during its training phase. The blue line represents the model's accuracy in predicting rewards, plotted against the number of training steps. Starting from near-zero accuracy, the model quickly improves its performance, reaching and maintaining an average high accuracy levels within the first 500 steps. The subsequent fluctuations shows the continuous job done by the fine-tuning as the model meets new data. The distinct step-like appearance of the graph is due to the model's virtual batch size of 2, which constrains the possible accuracy values to 0/2, 1/2, 2/2(corresponding to 0%, 50% and 100% accuracy). Notably, the overall trend of increasing accuracy, despite variations in the input predicates, suggests the model could be generalizing its learning which could be a key indicator of robust language understanding and generalization over predicates.

and future work might clarify and explain better what happens.

But we also find that after using the IRPO technique the model modifies its behaviour, improving its accuracy. In other words model seems to acquire some competence in this task by being fine-tuned with a double signal consisting of the DPO plus the NLL losses being considered. The first signal teaches the model to distinguish between the right and the wrong answer. The second signal pushes further up the correct answer in probability space. It is remarkable that the collection of the dataset for the second iteration proved to be quite a hard task since the model was performing well enough to give just a reduced amount of wrong answers, both in an absolute – i.e. for a given sentence the model returns 30 correct answers – and in a relative – i.e. the number of wrong answers is small: 2,3 – sense.

We assessed the model's performance on basic Common Sense [14] tasks to probe the effects of our fine-tuning. Interestingly, we found no change in performance across these tasks. This outcome is particularly noteworthy when we analyze how the different outcomes might have been speculated to have happened. A deterioration in performance could have suggested catastrophic forgetting, a common issue in neural networks where new learning replaces irremediably the previous knowledge. However, our use of Low-Rank Adaptation (LoRA) likely mitigated this effect by updating only task-specific parameters. The unchanged performance indicates that our fine-tuning enhanced the model's capabilities on our specific task without compromising its general language understanding. This result aligns with the versatility of large language models, capable of maintaining proficiency across multiple NLP tasks simultaneously, and suggests potential for developing specialized AI systems without sacrificing broader capabilities.

Another significant result derives from the fact that the subset of verbs used for fine-tuning differs from the verb subset used for testing. This means that we not only avoid using the same sentences from the training phase but also employ verbs that were not present during training, and yet we obtain performance improvements. This demonstrates some degree of generalization. While a thorough study would be needed to draw more accurate conclusions, we can still offer some speculations:

1. The model has generalized to understand which thought patterns are rewarded during the DPO (Direct Preference Optimization) phase. It now attempts to reproduce these patterns regardless of the specific sentence or verb used. This can be seen clearly from Figure 1, where the accuracy in rewarding the right choices improves over training steps, a key indicator that the model *learns* to discern between right and wrong answers.
2. Another speculation is that the logits have changed, and the behaviour is too random or at least too noisy to be taken into consideration alone. This means that the fact that the model performs better could be just associated with random guessing as the baseline models do, but with new logits.

We now briefly illustrate the impact of fine-tuning on the Mistral 7B model's performance in subcategorization tasks, presenting a case study focusing on the Italian verb ”scherzare” (to joke). This example, while not necessarily representative of all improvements, provides at least some insight into the changes in the model's reasoning process. The sentence under consideration is: ”Non si può scherzare con la realtà virtuale.” (One cannot joke with virtual reality.) The model was tasked with selecting the correct sense from a predefined set of options:

Human non fa o dice qualcosa sul serio (doesn't do or say something seriously)

Human fa ironia su|sopra|con [Abstract Entity] (makes irony about [Abstract Entity])

Human ha un comportamento leggero e divertito (has a light and amused behavior)

Human fa sul serio, non sottovaluta, prende sul serio ([Anything]) (takes seriously, doesn't underestimate [Anything])

Human si espone imprudentemente a un rischio (imprudently exposes oneself to a risk)

Human prende alla leggera qualcosa (takes something lightly)

Before fine-tuning, the model incorrectly classified the usage as sense 1, providing the following reasoning: ”Nella frase data, ’scherzare’ non è seguito da un oggetto concreto o astratto, quindi non si applicano i sensi 1, 2, 4, 5 e 6. Il contesto della frase indica che ’scherzare’ non è inteso come un comportamento leggero e divertito (senso 3), ma piuttosto come non fare o dire qualcosa in maniera irrispettosa o non seria (senso 1).” This reasoning demonstrates a lack of sensitivity to the syntactic structure and context of the sentence. Also, during the reasoning process, option 1 was inconsistently excluded but was later chosen as the final answer.

After fine-tuning, the model correctly identified sense 4 as the appropriate choice: ”Nella frase data, ’scherzare’ non è usato attivo, ma negativo, quindi non si può applicare direttamente i sensi 1, 3, 5 e 6, che si riferiscono a azioni o comportamenti specifici. Invece, il verbo ’si può’ implica un senso di non poter fare qualcosa, quindi il senso appropriato è quello di sottovalutare o prendere in giro qualcosa, che corrisponde al senso 4.” This *reasoning* shows multiple enhancements:

1. Recognition of the negative construction ”Non si può scherzare”
2. Consideration of the phrase ”con la realtà virtuale” as crucial context
3. More nuanced interpretation, considering multiple senses before making a decision

While this single example cannot be generalized to the model’s overall performance, it suggests that fine-tuning may have enhanced the model’s ability to parse complex syntactic structures and integrate contextual information in subcategorization tasks. Further comprehensive analysis across a wide range of verbs and constructions would be necessary to draw broader conclusions about the model’s improved capabilities as well as identifying new means to further enhance accuracy and performance.

Table 1
Comparison between various fine-tune methods

Model	Test Accuracy (%)
Iterative RPO	
Iteration 1	**75.6**
SFT	
PST CoT	65.6
Mistral baseline	
Zero-shot CoT	59.8

6. Conclusion and Future Work

In conclusion, we can say that small multilingual baseline models such as Mistral 7B perform poorly on semantic analysis of Italian sentences. We observe that the poor behavior is due to the model’s inability to discern the correct answer, either because it lacks the linguistic knowledge, therefore mostly resorting on random guesses, or because it follows an incorrect explanation for the answer is about to give. However, our research also demonstrates that the model can be significantly improved using IRPO techniques without affecting the baseline performance on common sense and reasoning tasks. Notably, we observe the ability to generalize across predicates, likely due to underlying linguistic skills, though further investigation is needed to fully understand this phenomenon.

The production of small open language models is rapidly evolving, approaching the level of huge close models which were available on the cloud a couple of years ago. At present, Italian monolingual models have room for improvement in terms of performance levels, [3] while multilingual models, e.g. the recently released `Gemma 2`[15], show increasing proficiency in our language, probably due to transfer learning effects. Our research shows the potential of leveraging such models in combination with high-quality lexical resources to develop a new class of task-specific models for the Italian language. These models, while small in scale, are expected to exhibit remarkable proficiency in executing complex analytical tasks, such as those related to verbs.

With this in mind, our future work is aimed, on the one hand at enriching lexicographic resources and refining the ways to obtain training material from them, and on the other hand at continuously evaluating the improvements brought about by the progress of general-purpose open models.

One promising application is the use of a verbal subcategorization and frame extraction system to extract content from specialist documents, such as legal [16] or medical texts [17]. Furthermore, the ability to analyze the complex argument structure of verbs has potential for use in language learning systems [18], e.g. providing support for immigrants to learn Italian affordably.

Finally, we made our fine-tuned model publicly available on huggingface[4] along with a visual report on wandb.[5]

[3]See for instance Hugging Face’s INVALSI Leaderboard, `https://huggingface.co/spaces/Crisp-Unimib/INVALSIbenchmark`
[4]https://huggingface.co/theGhoul21/srl-base-irpo-080524-16bit-v0.3-lighning-ai-6000
[5]https://shorturl.at/4jmPq

References

[1] K. K. Schuler, Verbnet: A broad-coverage, comprehensive verb lexicon, Ph.D. thesis, University of Pennsylvania (2005).

[2] C. F. Baker, C. J. Fillmore, J. B. Lowe, The berkeley framenet project, Proceedings of the 17th international conference on Computational linguistics-Volume 1 (1998) 86–90.

[3] M. Palmer, D. Gildea, P. Kingsbury, The proposition bank: An annotated corpus of semantic roles, in: Computational Linguistics, volume 31, MIT Press, 2005, pp. 71–106.

[4] L. Shi, R. Mihalcea, Putting pieces together: combining framenet, verbnet and wordnet for robust semantic parsing, in: International Conference on Intelligent Text Processing and Computational Linguistics, Springer, Mexico City, 2005, pp. 100–111.

[5] A.-M. Giuglea, A. Moschitti, Semantic role labeling via framenet, verbnet and propbank, in: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Sydney, NSW, 2006, pp. 929–936.

[6] S. W. Brown, J. Bonn, G. Kazeminejad, A. Zaenen, J. Pustejovsky, M. Palmer, Semantic representations for nlp using verbnet and the generative lexicon, Frontiers in Artificial Intelligence 5 (2022) 821697. doi:10.3389/frai.2022.821697.

[7] E. Jezek, B. Magnini, A. Feltracco, A. Bianchini, O. Popescu, T-pas: A resource of typed predicate argument structures for linguistic analysis and semantic processing, in: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), European Language Resources Association (ELRA), 2014, pp. 890–895.

[8] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, W. E. Sayed, Mistral 7b, 2023. URL: https://arxiv.org/abs/2310.06825. arXiv:2310.06825.

[9] R. Y. Pang, W. Yuan, K. Cho, H. He, S. Sukhbaatar, J. Weston, Iterative reasoning preference optimization, ArXiv abs/2404.19733 (2024). URL: https://api.semanticscholar.org/CorpusID:269457506.

[10] P. Hanks, Lexical analysis: Norms and exploitations, Mit Press, 2013.

[11] N. Wang, J. Li, Y. Meng, X. Sun, H. Qiu, Z. Wang, G. Wang, J. He, An MRC framework for semantic role labeling, in: N. Calzolari, C.-R. Huang, H. Kim, J. Pustejovsky, L. Wanner, K.-S. Choi, P.-M. Ryu, H.-H. Chen, L. Donatelli, H. Ji, S. Kurohashi, P. Paggio, N. Xue, S. Kim, Y. Hahm, Z. He, T. K. Lee, E. Santus, F. Bond, S.-H. Na (Eds.), Proceedings of the 29th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 2022, pp. 2188–2198. URL: https://aclanthology.org/2022.coling-1.191.

[12] R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, C. Finn, Direct preference optimization: Your language model is secretly a reward model, in: Thirty-seventh Conference on Neural Information Processing Systems, 2023. URL: https://openreview.net/forum?id=HPuSIXJaa9.

[13] J. E. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, W. Chen, Lora: Low-rank adaptation of large language models, ArXiv abs/2106.09685 (2021). URL: https://api.semanticscholar.org/CorpusID:235458009.

[14] L. Gao, J. Tow, B. Abbasi, S. Biderman, S. Black, A. DiPofi, C. Foster, L. Golding, J. Hsu, A. Le Noac'h, H. Li, K. McDonell, N. Muennighoff, C. Ociepa, J. Phang, L. Reynolds, H. Schoelkopf, A. Skowron, L. Sutawika, E. Tang, A. Thite, B. Wang, K. Wang, A. Zou, A framework for few-shot language model evaluation, 2023. URL: https://zenodo.org/records/10256836. doi:10.5281/zenodo.10256836.

[15] G. T. et al., Gemma: Open models based on gemini research and technology, 2024. URL: https://arxiv.org/abs/2403.08295. arXiv:2403.08295.

[16] S. Hassani, Enhancing legal compliance and regulation analysis with large language models, 2024. URL: https://arxiv.org/abs/2404.17522. arXiv:2404.17522.

[17] U. Mumtaz, A. Ahmed, S. Mumtaz, Llms-healthcare : Current applications and challenges of large language models in various medical specialties, 2024. URL: https://arxiv.org/abs/2311.12882. arXiv:2311.12882.

[18] N. Haristiani, Artificial intelligence (ai) chatbot as language learning medium: An inquiry, Journal of Physics: Conference Series 1387 (2019) 012020. URL: https://dx.doi.org/10.1088/1742-6596/1387/1/012020. doi:10.1088/1742-6596/1387/1/012020.

A. A complete example

A.1. Prompt example

This is an example of a prompt. The predicate is "allontanare" that in English can be translated based on the sense with expel, put at distance or also go away from a place. Another meaning is leaving and also repel or keep at distance something or someone. In this case we ask the model to understand what is the right sense to select among the 5 (in this case) possible choices. This is the sentence translated in English:

Those islanders, jealous of the small properties they had acquired at a very hard cost, intended with extreme penalties to remove the danger of seeing them taken away from them.

```
Frase da analizzare: "Quegli isolani ,
gelosi delle piccole proprietà, a prezzo
sudatissimo procurate , intendevano con
estreme pene allontanare il pericolo di
vedersele sottratte . "

Task: Ragionando sempre in italiano,
determina il significato del verbo
"allontanare" in base al contesto della
frase. Considera i seguenti sensi:

1:
Forma: [Human1] allontanare [Human2] da
[Location] | da [Human Group]
Significato: [Human1] espelle, caccia
[Human2] da [Location] | da [Human Group]
2:
Forma: [Human] allontanarsi (da [Physical
Entity] | da [Human Group])
Significato: [Human] fa in modo di essere
lontano (da [Physical Entity] | da [Human
Group])
3:
Forma: [Human] allontanarsi da [Location]
Significato: [Human] si assenta da [Location]
4:
Forma: [Human] allontanare [Abstract Entity]
Significato: [Human] rimuove, respinge, tiene
lontano [Abstract Entity]
5:
Forma: [Eventuality] allontanarsi
Significato: [Eventuality] si distanzia,
diventa difficilmente raggiungibile

Domanda: Quale dei sensi sopra elencati
meglio descrive l'uso del verbo
"allontanare" nella frase data? Spiega il
tuo ragionamento passo per passo prima di
rispondere e quanto sei pronto a rispondere
usa il formato
"Risposta finale: 1,2,3,4,5 ecc".
Soluzione:
```

A.2. The right answer

This is one of the example responses the baseline model gives and that is correct and will be used in the training as *winning*.

```
Prima di procedere con l'analisi,
letteralmente, la frase significa
"those islanders, jealous of their little
properties, obtained at a great price,
intended with extreme pain to keep the danger
of having them taken away at a distance."
Given the context, we can deduce that the
"danger" referred to in the phrase
is a potential threat that may cause
the "islanders" to lose their "little
properties."
Therefore, in order to protect these
possessions, the islanders are taking
deliberate actions to prevent the danger from
approaching.
Given this context, option 1 is the least
likely to fit, as there is no
explicit indication that anyone is being
expelled or cast out. Option 5
also seems unlikely, as the islanders are
acting intentionally to keep
the danger at bay, not distancing themselves
from it.
Options 2, 3, and 4 all have potential, but
let's consider each in turn.
In option 2, the verb "allontanarsi" does not
appear, and the rest of
the sentence suggests a collective rather
than individual action, which
makes it less likely. Option 3 seems
plausible at first, but it doesn't
fully capture the meaning of the text. While
the islanders may be
physically separating themselves from the
danger, the primary intent
of their actions is to keep the danger away
from their properties,
not just to be far from it.
On the other hand, option 4 seems to fit the
context best. The islanders
are taking deliberate steps to remove or keep
the danger at a distance
from their possessions, effectively
"removing" or "keeping at a distance"
the potential threat to their property.
Risposta finale: 4.
```

A.3. The wrong answer

On the other hand another response given by the model to the same prompt says that the islanders are pushing someone away but is not explicitly stated in the sentence. But the islanders are removing the danger which is sense 4, that is keeping at distance an Abstract Entity. This answer is obviously wrong

Considerando il contesto della frase, "quegli isolani intendevano con
estreme pene allontanare il pericolo di vederle sottratte," il senso
secondo (3) "Forma: [Human] allontanarsi da [Location]" e (4) "Forma:
[Human] allontanare [Abstract Entity]" non sembrano appropriati. Infatti, il primo senso implica un'azione volontaria e attiva di uno soggetto verso un'altra persona, mentre nel contesto di questa frase si intende un'azione collettiva compiuta dai "quegli isolani"
per proteggere "le piccole proprietà" dalle minacce esterne.
Il senso quinto invece non è applicabile in quanto non si tratta di una entità astratta a essere allontanata, bensì di un pericolo da mantenere lontano.
Perciò, il senso appropriato del verbo "allontanare" in questa frase è il primo
(1): "Forma: [Human1] allontanare [Human2] da [Location] | da [Human Group]".
In questo caso, "quegli isolani" stanno cacciando o espellendo un pericolo (che è non specificato nella frase) lontano dai loro piccoli propriety.
Risposta finale: 1.