Evaluating Models, Prompting Strategies, and Task Formats: a Case Study on the MACID Challenge

Matteo Rinaldi[1,*], Rossella Varvara[1], Lorenzo Gregori[2] and Andrea Amelio Ravelli[3,4]

[1]University of Turin, Corso Svizzera 185, 10149 Torino, Italy
[2]University of Florence, Via della Pergola 60, 50126 Firenze, Italy
[3]University of Trento, Via Giuseppe Verdi 26, 38122 Trento, Italy
[4]University of Bologna, Via Cartoleria 5, 40124 Bologna, Italy

Abstract

In this study, we test the ability of 8 Large Language Models to discriminate closely related action concepts, based on textual descriptions or on video representations. Our aim is to understand if these models can handle the fine-grained action understanding that humans perform with ease, particularly when there are cases of action-predicate mismatches, i.e., the same verb may describe different actions, or different verbs may refer to the same action. We experiment on the MACID dataset, a dataset of actions representing "pushing" events and manually annotated for action IDs taken from the IMAGACT ontology. We evaluate how prompt complexity and task formats influence models' performance. Particularly, we test three different prompts with or without examples, two task formats (binary or multiple choice task), and two modalities (textual or visual). Results indicate that the binary task is not easier than the multiple-choice one, and that few-shot prompting generally improves models' accuracy. Moreover, LLMs perform better when helped by lexical cues: accuracy increases when actions are expressed by different verbs, whereas it is lower when actions are expressed by the same verb.

Keywords

large language models, action concept understanding, prompting strategies, task definition

1. Introduction

Understanding human action is a cornerstone of both linguistic and perceptual intelligence. The close interdependence between language and vision in human cognition is suggested by the Mirror System Hypothesis [1], which considers language as not merely symbolic but grounded in sensorimotor experience. This cognitive grounding implies that effective language understanding, especially of action-related expressions, requires grasping subtle distinctions between closely related actions. Recent advances in large language models (LLMs) and the emergence of multimodal LLMs, which are capable of jointly processing textual and visual inputs, allow the integration of perceptual and linguistic reasoning in artificial models. However, it remains unclear to what extent these models can handle the fine-grained action understanding that humans perform with ease, particularly when linguistic descriptions are ambiguous or semantically close. To address this gap, we investigate the performance of both textual and multimodal LLMs on the MACID dataset [2], a benchmark specifically designed to evaluate the capacity of models to distinguish between subtly different human actions described using similar or identical linguistic expressions. The MACID dataset provides both natural language descriptions and corresponding video clips of the actions, enabling an evaluation of how visual grounding can support or enhance linguistic disambiguation. In this paper, we aim to test the strengths and limitations of current LLMs in grounded language understanding by analyzing the ability of LLMs to resolve action ambiguities from linguistic or visual input. We experiment considering 8 LLMs, two task formats, three prompts of increasing complexity, and two modalities (visual or textual). We compare models' results to random baselines, and we evaluate the role of the lexical component in the disambiguation of actions.

2. Related work

2.1. Action concepts definition

Following the conceptual framework at the basis of the IMAGACT Ontology of Actions [3], we define an *action concept* as a cognitively grounded and language-independent representation of a physical action involving an agent modifying the world. Action concepts are generalizable across contexts, i.e., may apply to different agents and objects, and they are encoded in the IMAGACT Ontology through prototypical scenes, in the form of short videos, which visually disambiguate verb meanings. The

CLiC-it 2025: Eleventh Italian Conference on Computational Linguistics, September 24 — 26, 2025, Cagliari, Italy
*Corresponding author.
matteo.rinaldi@unito.it (M. Rinaldi); rossella.varvara@unito.it (R. Varvara); lorenzo.gregori@unifi.it (L. Gregori); andreaamelio.ravelli@unitn.it (A. A. Ravelli)
0009-0004-7488-8855 (M. Rinaldi); 0000-0001-9957-2807 (R. Varvara); 0000-0001-9208-2311 (L. Gregori); 0000-0002-0232-8881 (A. A. Ravelli)

relation between verbs and action concepts is not one-to-one: a single verb may express different concepts, and a concept may be lexicalized by multiple verbs. IMAGACT's multimodal approach supports cross-linguistic comparison and enables accurate mapping between verbs in multiple languages and their underlying event structures, independent of syntactic realization or argument structure. These form-meaning mismatches make action concepts foundational for modeling verb semantics in both theoretical and computational settings.

2.2. LLM benchmarking

Large Language Models are usually introduced to the community by showcasing their very high performance on classic benchmarks. They are very good at solving complex math problems, writing and debugging code, or answering multiple-choice questions about common knowledge. However, this kind of evaluation does not tell the full story. When LLMs are tested on more realistic tasks, i.e., closer to what a normal person might do, they often lose their *super-human* performance. These models still struggle with tasks that truly require human-like understanding, such as subtle semantic variations, pragmatic understanding, and so on. So, even if they do very well on traditional benchmarks, their performance in real-life or more *everyday human life* tasks is still limited.

Moreover, most of the research and effort in this field is on the English language. The CALAMITA benchmark [4] represents the first of its kind as an Italian-focused collection of tasks that really pose a challenge for commonsense, factual, and linguistic knowledge in Italian.

3. Experimental Setting

3.1. Data

The data used in this study is taken from the CALAMITA benchmark [4], specifically from the MACID challenge [2]. This dataset is based on a portion of the LSMDC dataset [5], a collection of short video clips extracted from movies along with transcriptions of English DVS (descriptive video services) for visually impaired people. The LSMDC dataset is the result of the merging of two previous datasets, both built upon DVS from movies: the Max Planck Institute für Informatik Movie Description Dataset (MPII-MD) [6], and the Montreal Video Annotation Dataset (M-VAD) [7]. The textual captions were manually translated into Italian and modified to depict the action in the corresponding video and to avoid vague references (e.g., pronouns substituted with common nouns).

The MACID dataset includes video-caption pairs restricted to a set of similar actions, i.e. to the variation of actions and action verbs linked to "pushing" events. This choice was made to define a challenging task, in which subtle semantic differences occur among the different items. Data have been manually filtered and annotated [8] using the action conceptualization derived from the IMAGACT Multilingual and Multimodal Ontology of Actions [3]. IMAGACT is a multimodal and multilingual ontology of actions that provides a fine-grained categorization of action concepts, each represented by one or more visual prototypes in the form of recorded videos and 3D animations. IMAGACT currently contains 1,010 scenes that encompass the action concepts most commonly referred to in everyday language usage. Scenes belonging to the same action concept are grouped together and labeled with a unique identification number. The categorization of action concepts proposed in the theoretical framework behind IMAGACT has been validated in a series of experiments with a high inter-annotator agreement [9], confirming that the theoretical framework can be considered well-founded and reproducible.

3.2. Task formats

Models are evaluated on two distinct versions of the MACID dataset. Initially, models are assessed on an intruder detection task in sets of four sentences: three sentences are related to the same action concept while one is related to a different action concept. The goal of the model is to correctly identify the intruder sentence within each set, that is, the only one referring to an action concept different than the remaining three.

The second experiment is performed on the binarized version of the MACID dataset: models were required to compare sentence pairs and classify them as either "different" or "equivalent" with respect to the action concept expressed by the sentence.

3.2.1. Multiple choice

The dataset in the original MACID challenge [2] was structured on groups of 4 captions, three of which were annotated as belonging to the same action concept, and one describing a different action type. Each entry in the dataset is structured as follows:

- *id*: the quadruple id;
- *s1-4*: the 4 caption sentences describing the actions;
- *v1-4:* the reference ID of the 4 videos depicting the actions;
- *intruder*: the number (1-4) of the sentence (and video) which is the intruder in the group.

Video files are provided in an additional folder, named with a unique reference ID.

An example of the textual data follows.

QUADRUPLE_1

(1) I due ragazzi spingono il carrello verso la colonna (*The two boys push the cart toward the column*)
[action id: 65431186]

(2) La donna spinge la signora anziana sulla sedia a rotelle (*The woman pushes the elderly lady in the wheelchair*)
[action id: 65431186]

(3) L'uomo spinge a terra l'aggressore (*The man pushes the attacker to the ground*)
[action id: 18ad2fa9]

(4) L'infermiere spinge la barella (*The nurse pushes the gurney*)
[action id: 65431186]

3.2.2. Binary choice

In order to verify the impact of the task format on this challenge, we converted the dataset (as well as the task) into a binary format. This second dataset consists of video-caption pairs, together with their action concept IDs and the information about whether they correspond to the same action type or not. We kept the information about the quadruple ID to allow comparison between the results from the two formats. The columns in the new version of the dataset describe the following information:

- *id*: the quadruple id;
- *s1-2*: the 2 caption sentences describing the actions;
- *v1-2:* the reference IDs of the 2 videos depicting the actions;
- *id1-id2*: the action concept IDs of the 2 actions;
- *different*: information about the actions being different (1) or the same (0).

3.3. Models

For this experiment, we tested a bunch of textual models: five small models with 7/8/9 billion parameters (Llama3.1, Qwen2.5, Aya-expanse, Mistral, Minerva, Gemma2), one medium native-Italian model with 14 billion parameters (Velvet), and one big model with 72 billion parameters (Qwen2.5).

4. Prompting strategies

In both scenarios (multiple or binary choice), we tested three prompts, built with incremental information. The first prompt (SHORT) is the same proposed for the original MACID Challenge, and it is a baseline with just the necessary information to execute the task. The second prompt (MEDIUM) adds to the first more details about what an action concept is, and what are the main features which discriminate between close but different actions. The third prompt (LONG) elaborates more on the theoretical distinction between actions and is enriched with some explanation about the possible mismatch between actions and verbs. Finally, we added to the experimental setting a version of the task without any explanation (NONE), but with only some examples. All prompts were formulated in Italian to assess both the models' sentence processing capabilities and their ability to correctly interpret instructions given in the Italian language. All prompts are reported in the Appendix A.

4.1. Zero or few-shot settings

The empirical investigation with different prompting strategies aimed at finding the optimal balance between instructions given in a concise form and instructions given using a long and verbose language. This exploration involved developing three distinct prompts for each dataset variant, alongside an additional experiment utilizing few-shot examples without explicit instructions.

To expand the analysis on how the instruction given in the prompt influences the outcomes, each prompt was tested under both zero-shot and few-shot conditions. Five examples were selected from the quadruple dataset and four from the paired dataset, with consistent example sets maintained throughout the evaluation process. The selection of five examples from the quadruple dataset was strategically designed to encompass all possible verb relationship combinations: one example featuring four distinct verbs, one with three different verbs, one containing two different and two identical verbs, one with verbs paired identically, and one where all verbs were identical.

4.2. Textual and visual settings

In order to test the models on the different settings proposed in the MACID's experiments, we wrote a Python script that interrogates an OpenAI API compatible backend to perform interrogation and evaluation of the models. The script loads the data from JSONL files and formulates the different complete prompts for each datapoint. To evaluate the results, the scripts only consider the first sampled token and check if it corresponds to the expected outcome. For the experiment on quadruples, only the first character of the first token is considered and checked against the number identifying the intruder sentence. In the experiment of couples, considered that the model was asked to answer either "yes" (*si*) or "no" (*no*), the first sampled token was converted in lower case and accents were removed, so that it was possible to check it regardless of the case or the use of the accent on the word

sì, required in formally correct Italian but that may be omitted without changing the sentence's meaning even by native speakers. As a backend, we employed vLLM with Flash Attention 2.7 for optimal performance for all the 7B, 8B and 14B models. Qwen 2.5 72b was instead accessed using the "OpenRouter" API and loaded with BF16 weights. All the models were set to a temperature of 0.0 and a random seed of "27" in order to obtain reproducible results. All the results were then saved in a SQLite database for easy access.[1]

We decided to purposely opt for a strict evaluation strategy: answers where the model wrote any kind of text before the actual task's answer - such as chattering, boilerplate text, reasoning traces, or unwanted answer's formatting - were automatically discarded by the evaluation script, that expected the correct answer to be in the very first characters of the model's response. This decision is motivated by the fact that we also wanted to test the models' capabilities to strictly adhere to the given instructions: a model that talks too much or return the answer in an unwanted format is a model that may pose problems in production scenario, such as higher costs, due to the generation of more tokens, or the need to add post-processing strategies.

5. Results

In this section, we discuss the results obtained across all the experimental scenarios (i.e., prompting strategies, zero/few-shot, multiple/binary choice). On both task formats, we defined a majority class baseline. The baseline accuracy for the multiple choice task is 28% , while for the binary choice task it is 50%.

5.1. Results with textual LLMs

Figure 1 reports the performance of the models tested in both multiple-choice (1a) and binary-choice (1b) tasks. Before illustrating the results, we present an evaluation of the ability of the models to follow the instructions and to provide the answer in the required format. Indeed, we forced the model to reply with only 4 tokens, since we expected a yes/no answer for the binary task and a number to identify the intruder sentence in the multiple-choice task. The desired output format has been unambiguously specified in the prompts (see Appendix A), although we decided not to be strict in accepting answers: upper/lower case, accents, or additional spacing, have been tolerated whenever the "yes/no" or "1/2/3/4" strings were present in the answer. We didn't use any additional tool to constrain the output (e.g, Guidance[2], Outline[3]), because the requested output format is straightforward and we considered a good adherence to it as part of the task. Restricting the amount of output tokens to 4 also allowed for a great saving of resources, given the high computational costs of autoregressive generation.

Some models were not able to perfectly adhere to the instructions, but this behavior seems related to some task formats. *Aya-expanse-8b* does not follow the required format with all three prompts when tested for binary response without examples. *Gemma-2-9b* provides unacceptable responses for all the binary task's conditions.[4] *Minerva-7B-instruct-v1.0*, with no difference between prompts and binary/multiple choice tasks, does not adhere in the zero-shot setting, with the exception of the short prompt in the binary task.

Binary choice task Among the small models (ranging between 7 and 14 billion parameters), *llama-3.1-8b-instruct* reaches the best results, with a .696 accuracy when instructed with the long prompt in a few-shot setting. This model reaches high accuracy (.689) even with the short prompt with examples and with the examples alone, showing generally a preference for the few-shot setting with respect to the zero one (with a .133 difference in accuracy between the few and zero-shot setting with the long prompt, Table 1).

Qwen-2.5-72b reaches the highest accuracy (.725) among all models, with the long prompt and the few-shot setting. However, despite the huge difference in parameters, it is outperformed in short_zero setting by *Llama-3.1-8b*. As noted above, some models (i.e., Minerva-7b and aya-expanse-8b) do not provide satisfying replies in some conditions (marked as ND in Table 1).

In general, the few-shot setting improves the results in the binary task, even if in some cases the difference is small.

With regard to the prompt type, 5 models out of 7 show a preference for the long prompt. *Aya-expanse-8b* does slightly better with the medium prompt (.647) with respect to the detailed prompt (.640), whereas *Velvet-14B* achieves the same accuracy with both (.507).

Native Italian models do not perform better than the others: the results from *Velvet-13b* are close to chance, whereas *Minerva-7b* achieves better in the long-few shot setting.

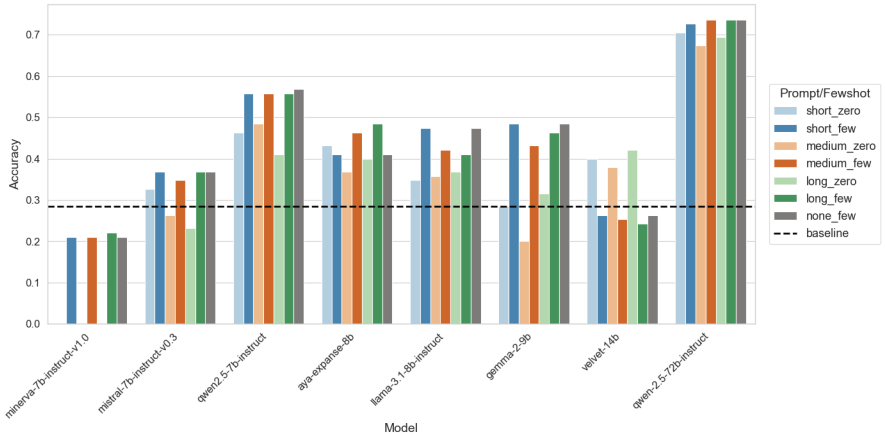
We additionally analyze the impact of the lexical component on models' performance, i.e., we look at if and how models are facilitated when actions are expressed by different verbs (Table 5, Appendix B) and when they are expressed by the same one (Table 4, Appendix B). Most models achieve higher accuracy when actions are

[1] All data and scripts are available at https://github.com/mrinaldi97/MACID/

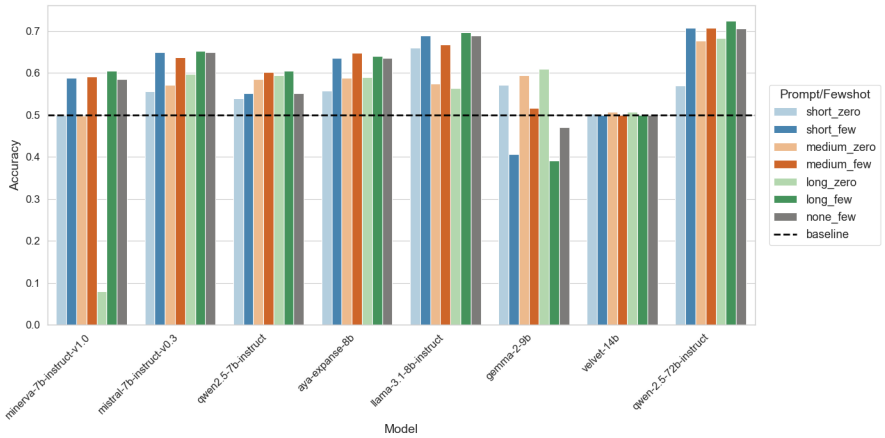
[2] https://github.com/guidance-ai/guidance

[3] https://github.com/dottxt-ai/outlines

[4] Given this behavior, we excluded Gemma-2-9b from the summary tables reported in Appendix.



(a) Multiple Choice task format



(b) Binary choice task format

Figure 1: Comparison of models in all the experimental scenarios, both in the multiple choice (1a) and in the binary choice (1b) task configurations.

expressed by different verbs: it is easier to discriminate if two sentences express the same action if their lexical description is different as well. When the verbs are equal, accuracy decreases. This difference is smoother when examples are added in the prompts, and it increases with the short prompt. A notable exception is given by *llama-3.1-8b-instruct*, which achieves higher accuracy for actions expressed by the same verbs rather than with different verbs (reaching a value of .933 in the long-zero format). When looking in more detail at its behavior, we note that this happens with the two most detailed prompts, and we hypothesize that it may be due to specification that there is no one-to-one matching between action concepts and verbs included in these prompts.

Multiple choice task Among the small models, *qwen2.5-7b* reaches the best results, with a .568 accuracy when instructed with the examples. However, differently from the binary task, the gap with the larger model (*qwen-2.5-72b*) is notable, with the latter performing very well among all conditions and reaching an accuracy of 0.737 in three of them (few-shot with medium, long, and no prompt). Even if it has been noted frequently that LLMs do not perform well with multiple-choice tasks, in this challenge, they do better than in the binary choice one, considering the random baseline for each task (Table 2.

As noted for the binary task, providing a few examples increases accuracy. Exceptions, however, are found for the short prompt: *velvet-14b* and *aya-expanse-8b* have a slightly higher accuracy with the zero-shot setting with respect to the few-shot. The zero/few shot setting also

Model	short zero	short few	medium zero	medium few	long zero	long few	none few	average
minerva-7b-instruct-v1.0	0.500	0.588	0.498	0.591	0.079	0.605	0.584	0.492
mistral-7b-instruct-v0.3	0.556	0.649	0.572	0.637	0.596	0.653	0.649	0.616
qwen2.5-7b-instruct	0.539	0.551	0.584	0.602	0.595	0.605	0.551	0.575
aya-expanse-8b	0.558	0.635	0.588	0.647	0.589	0.640	0.635	0.613
llama-3.1-8b-instruct	**0.660**	0.689	0.574	0.667	0.563	0.696	0.689	0.648
gemma-2-9b	0.572	0.406	0.595	0.516	0.609	0.391	0.470	0.508
velvet-14b	0.502	0.500	0.507	0.500	0.507	0.500	0.500	0.502
qwen-2.5-72b-instruct	0.570	**0.707**	**0.677**	**0.707**	**0.682**	**0.725**	**0.705**	**0.682**
BASELINE	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500

Table 1
Models accuracy in the Binary Choice task.

Model	short zero	short few	medium zero	medium few	long zero	long few	none few	average
minerva-7b-instruct-v1.0	0.000	0.211	0.000	0.211	0.000	0.221	0.211	0.122
mistral-7b-instruct-v0.3	0.326	0.368	0.263	0.347	0.232	0.368	0.368	0.325
qwen2.5-7b-instruct	0.463	0.558	0.484	0.558	0.411	0.558	0.568	0.514
aya-expanse-8b	0.432	0.411	0.368	0.463	0.400	0.484	0.411	0.424
llama-3.1-8b-instruct	0.347	0.474	0.358	0.421	0.368	0.411	0.474	0.408
gemma-2-9b	0.284	0.484	0.200	0.432	0.316	0.463	0.484	0.380
velvet-14b	0.400	0.263	0.379	0.253	0.421	0.242	0.263	0.317
qwen-2.5-72b-instruct	**0.705**	**0.726**	**0.674**	**0.737**	**0.695**	**0.737**	**0.737**	**0.716**
BASELINE	0.280	0.280	0.280	0.280	0.280	0.280	0.280	0.280

Table 2
Models accuracy in the Multiple Choice task.

has an influence also in the ability of *Minerva-7b* to comply with the required output format: when provided with examples, it follows the instructions, whereas it does not in the zero-shot prompt.

Contrary to what we observed above for the binary task, the prompt type does not widely influence the results: accuracy values for most models (*minerva-7b, mistral-7b, qwen2.5-7b, llama-3.1-8b, qwen-2.5-72b*) are equal among different prompts.

As for the binary task, the verb used to describe the intruder has an impact: if it is the same as (at least one of) the other sentences, models' performance drops, even if less strongly (Table 6 and 7 in Appendix B).

5.2. Results with visual LLMs

The MACID dataset includes all the original videos referred to by the sentences. This setting enabled us to conduct an exploratory experiment with multimodal models, particularly those capable of processing video inputs. At the time of writing, video models are in their early developmental stages. A great effort is going on to understand optimal methods for integrating video information into language models, as video data presents challenges for transformer architectures due to the quadratic computational cost of self-attention over long sequences. Moreover, different research groups are experimenting with different architectural choices to ensure an effective alignment between video and language latent spaces [10]. We conducted experiments with two state-of-the-art video models: Qwen 2.5 VL 8B [11] and VideoLLama3 7B [12]. The models were executed on a local machine using configurations recommended in the official documentation. Both Qwen and VideoLLama utilize Hugging Face's "transformers" library, which includes the necessary code for running these video models. Both models handle videos of arbitrary resolution sampled at user-defined framerates. To keep memory usage manageable, we resized the original videos to 360x288 resolution. While this resolution is lower than the original files, often in FullHD (1920x1080) or PAL DVD (720x576) format, it remains perfectly intelligible to human viewers, being comparable to VideoCD (352x288) and VHS tape quality (240 horizontal TV lines). The framerate was set to 8fps because we decided to avoid very low framerates, given that video samples are brief (<4s) and consistently represent live action. Following the text-only experiments, we selected the best-performing prompt on average and adapted it for video model testing. Specifically, we modified the medium prompt to accommodate the video experiment, substituting sentences with video clips. Due to memory constraints, we executed the experiment exclusively on the binary task. Neither Qwen VL nor VideoLLama successfully handled the task: both models always returned "No" for every tested video pair. Interestingly, Qwen VL also provided brief video descrip-

tions. We speculate that the poor performance of video models on this task relates to difficulties in coherently processing temporal sequences and performing cross-domain inferences between visual and textual features. Moreover, the prompt being written in Italian and the presentation of two videos simultaneously, rather than the single-video setting usually employed during pre-training, further deviated the experimental conditions from the training distribution, substantially increasing task complexity. Testing multimodal and, in particular, video models poses significant challenges, and we believe that the Macid task can become a useful task to assess the models' abilities to correctly identify complex actions. For this reason, we leave to future work a more extensive experimentation with video models, including prompt/-formulation modifications, testing new models, as well as trying fine-tuning operations.

5.3. Discussion

Model	Average error rate
minerva-7b-instruct-v1.0	0.166
mistral-7b-instruct-v0.3	0.0
qwen2.5-7b-instruct	0.0
aya-expanse-8b	0.306
llama-3.1-8b-instruct	0.0
gemma-2-9b	0.867
velvet-14B	0.0
qwen-2.5-72b-instruct	0.0

Table 3
Average error rate for each model, grouped and averaged for all tasks.

Table 3 reports the average values of unacceptable responses per model, in each task, i.e. responses where the models did not adhere to the requested output format. As already stated, beside the objective of testing the ability of LLMs to interpret and discriminate descriptions of physical actions, we also want them to show their ability to follow the instructions given to them. One of the main problem we faced with our experiments is that responses from models tend to be overly verbose, as models need to explain their choices every time. While this may be considered a useful and interesting behavior in *chat* models, it is definitely not ideal in *instruct* models, as those tested in our experiments. As it is specified in all our prompts, we explicitly ask to answer with the id of the intruder for the multiple-choice and with "sì" or "no" (yes or no) for the binary-choice task (see Appendix A), thus the request is clear. Nevertheless, sometimes models tend to elude the requested response format (i.e., the answer does not start with a valid id number for the multiple-choice task, or it does not start with "sì/no" for the binary-choice task), while others apply absolutely unnecessary markup (e.g., *aya-expanse-8b*). Our evaluation framework (i.e., string matching) might appear at first glance to be simplistic, lazy, and excessively punitive for the models. As we already mentioned in Section 5.1, we could have used specific libraries to parse the responses in search of the correct result, but the point is that, given these models' reputation as "intelligent" (as promoted by the developers), one expects these models to be able to follow very simple instructions, regardless of their ability to effectively solve a task. Even in few-shot scenarios, where the requested answer format it is more than explicit, some models consistently fail in following the instructions. Models with *super-human* abilities might not need to be hand-guided.

6. Conclusions

This study evaluates LLMs on the action concepts discrimination task: we present the results for 7 LLMs evaluated on the MACID dataset.

Results show a wide variation in models' performances, depending on the model type, the number of model parameters, the prompt used, and the task format.

Qwen-2.5-72b obtained the highest average accuracy both on the binary and the multiple-choice task, confirming that the number of parameters is a core factor in this type of semantically complex task.

Italian models (Minerva and Velvet) perform poorly in both task formats. This is an unexpected result, considering the task requires fine-grained semantic abilities.

Among 7B/8B models, top results are achieved by Qwen-2.5, in multiple-choice format (acc. 0.568), and Llama-3.1 in binary format (acc. 0.696). The latter obtains an accuracy comparable with Qwen-2.5-72b (0.725), despite the difference in the number of parameters.

On average, few-shot prompting works better than zero-shot, both in binary and in multiple-choice task formats. In general, we don't find strong performance differences among the three versions of the task description in the prompt (SHORT, MEDIUM, and LONG), while there is a consistent accuracy improvement with the few-shot prompting. Even the few-shot without task description (*none_few*) has a good accuracy on the top models.

Finally, the lexical components have a strong influence on models' behavior in this task: the accuracy varies a lot if the two sentences use the same verb or different verbs (in the binary task) or if the intruder has the same verb as the other sentences or not (in the multiple-choice task). The accuracy gap between these two cases is huge with Qwen, which seems to be more sensitive to lexical differences than Llama. For example, Qwen-2.5-72b on a binary task reaches 0.975 accuracy with different verbs and 0.579 with the same verb.

Further experiments need to be done with video LLMs, which did not provide satisfactory results in this first experimentation.

References

[1] M. Arbib, G. Rizzolatti, Neural expectations: A possible evolutionary path from manual skills to language, Communication and Cognition 29 (1996) 393–424.

[2] A. A. Ravelli, R. Varvara, L. Gregori, MACID - multimodal ACtion IDentification: A CALAMITA challenge, in: F. Dell'Orletta, A. Lenci, S. Montemagni, R. Sprugnoli (Eds.), Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024), CEUR Workshop Proceedings, Pisa, Italy, 2024, pp. 1234–1238. URL: https://aclanthology.org/2024.clicit-1.137/.

[3] M. Moneglia, S. W. Brown, F. Frontini, G. Gagliardi, F. Khan, M. Monachini, A. Panunzi, et al., The imagact visual ontology. an extendable multilingual infrastructure for the representation of lexical encoding of action, in: Proceedings of the Ninth International Conference on Language Resources and Evaluation–LREC'14, European Language Resources Association (ELRA), 2014, pp. 3425–3432.

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

[5] A. Rohrbach, A. Torabi, M. Rohrbach, N. Tandon, C. Pal, H. Larochelle, A. Courville, B. Schiele, Movie description, International Journal of Computer Vision 123 (2017) 94–120.

[6] A. Rohrbach, M. Rohrbach, N. Tandon, B. Schiele, A dataset for movie description, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3202–3212.

[7] A. Torabi, C. Pal, H. Larochelle, A. Courville, Using descriptive video services to create a large data source for video annotation research, arXiv preprint arXiv:1503.01070 (2015).

[8] A. A. Ravelli, Annotation of linguistically derived action concepts in computer vision datasets, Ph.D. thesis, University of Florence, 2020.

[9] G. Gagliardi, Rappresentazione dei concetti azionali attraverso prototipi e accordo nella categorizzazione dei verbi generali. una validazione statistica, in: Proceedings of the First Italian Conference on Computational Linguistics–CLiC-it, 2014, pp. 180–185.

[10] K. Y. Y. Nakamizo, Act-ChatGPT: Introducing Action Features into Multi-modal Large Language Models for Video Understanding, Pattern Recognition(ICPR 2024) (2024).

[11] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, H. Zhong, Y. Zhu, M. Yang, Z. Li, J. Wan, P. Wang, W. Ding, Z. Fu, Y. Xu, J. Ye, X. Zhang, T. Xie, Z. Cheng, H. Zhang, Z. Yang, H. Xu, J. Lin, Qwen2.5-vl technical report, 2025. URL: https://arxiv.org/abs/2502.13923. arXiv:2502.13923.

[12] B. Zhang, K. Li, Z. Cheng, Z. Hu, Y. Yuan, G. Chen, S. Leng, Y. Jiang, H. Zhang, X. Li, P. Jin, W. Zhang, F. Wang, L. Bing, D. Zhao, Videollama 3: Frontier multimodal foundation models for image and video understanding, 2025. URL: https://arxiv.org/abs/2501.13106. arXiv:2501.13106.

A. Prompts

Prompts used for the experiments in binary and multiple-choice tasks.

Binary task

Zero-shot prompts

Three variants have been used, with increasing description details.

1. In questo task ti verranno proposte coppie di frasi che descrivono azioni fisiche. Il tuo compito è di indicare se le seguenti coppie di frasi esprimono lo stesso concetto azionale oppure no. Rispondi 'Sì' se ritieni che entrambe le frasi si riferiscano allo stesso concetto azionale, rispondi 'No' se ritieni che descrivano due concetti azionali diversi.
2. In questo task ti verranno proposte coppie di frasi che descrivono azioni fisiche. Le azioni nelle coppie possono essere dello stesso tipo, ovvero possono rappresentare lo stesso concetto azionale, oppure essere di due tipi diversi. Il tuo compito è di indicare se le seguenti coppie di frasi esprimono lo stesso concetto azionale oppure no. Un concetto azionale è un'entità linguistico-cognitiva corrispondente a un pattern di modifiche del mondo compiute da un agente, ed è generalizzabile a vari oggetti (o azioni). Un concetto azionale può essere realizzato linguisticamente con più verbi e, viceversa, un verbo può rappresentare più concetti azionali distinti. Rispondi 'Sì' se ritieni che entrambe le frasi si riferiscano allo stesso concetto azionale, rispondi 'No' se ritieni che descrivano due concetti azionali diversi.
3. In questo task ti verranno proposte coppie di frasi che descrivono azioni fisiche. Le azioni nelle coppie possono essere dello stesso tipo, ovvero possono rappresentare lo stesso concetto azionale, oppure essere di due tipi diversi. Il tuo compito è di indicare se le seguenti coppie di frasi esprimono lo stesso concetto azionale oppure no. Un concetto azionale è un'entità linguistico-cognitiva corrispondente a un pattern di modifiche del mondo compiute da un agente (umano, animale o macchina), ed è generalizzabile a vari oggetti (o azioni). Si tratta di una rappresentazione cognitiva di un evento o di un processo che coinvolge, prototipicamente, un agente (chi compie l'azione), un tema o paziente (sul quale si esercita l'azione) e, talvolta, uno strumento, un destinatario o una destinazione. Un concetto azionale è produttivo, ovvero può applicarsi a un'ampia varietà di oggetti e si presenta in contesti diversi. L'associazione tra concetto azionale e verbo che lo descrive non è un rapporto di tipo uno-a-uno. Infatti, un concetto azionale può essere realizzato linguisticamente con più verbi (ad es. 'spostare una

scatola' e 'spingere una scatola') e, viceversa, un verbo può rappresentare più concetti azionali distinti (ad es. 'aprire una porta' vs. 'aprire una noce'). Nell'individuare un concetto azionale, è importante concentrare l'attenzione su quali cambiamenti vengono compiuti dall'azione rappresentata, non sul verbo. Rispondi 'Sì' se ritieni che entrambe le frasi si riferiscano allo stesso concetto azionale, rispondi 'No' se ritieni che descrivano due concetti azionali diversi.

Few-shot prompts

Few-shot prompts are created by appending 4 examples to the three variants of zero-shot prompts; additionally, a fourth prompt with only examples and no description is provided. The following examples have been used.

1) I ragazzi spingono i carrelli lungo il binario del treno
2) La donna con gli occhiali da sole spinge l'anziana signora sulla sedia a rotelle
Risposta: Sì

1) L'uomo spinge una carriola nel cortile della fattoria mentre parla con la donna
2) Il veterinario spinge lo stantuffo della siringa
Risposta: No

1) La donna preme sul posacenere al centro del tavolo
2) Il ragazzo spinge le scope nel ripostiglio
Risposta: No

1) La donna sposta leggermente la tenda di perline
2) La donna spinge in alto il pannello di vetro
Risposta: Sì

Multiple-choiche task

Zero-shot prompts

1. In questo task ti verranno proposte quattro frasi che descrivono azioni fisiche. Tre di queste azioni sono dello stesso tipo, mentre una è di un tipo diverso. Individua la frase che descrive l'azione di tipo diverso. Esiste solo una risposta esatta, rispondi utilizzando esclusivamente il numero di riferimento della frase e nient'altro.
2. In questo task ti verranno proposte quattro frasi che descrivono azioni fisiche. Tre di queste azioni sono dello stesso tipo, ovvero rappresentano lo stesso concetto azionale, mentre una è di un tipo diverso. Un concetto azionale è un'entità linguistico-cognitiva corrispondente a un pattern di modifiche del mondo compiute da un agente, ed è generalizzabile a vari oggetti (o azioni). Un concetto azionale può essere realizzato linguisticamente con più verbi e, viceversa, un verbo può rappresentare più concetti azionali distinti. Tra le seguenti quattro frasi, individua la frase che descrive l'azione di tipo diverso dalle altre tre. Esiste solo una risposta esatta, rispondi utilizzando esclusivamente il numero di riferimento della frase e nient'altro.
3. In questo task ti verranno proposte quattro frasi che descrivono azioni fisiche. Tre di queste azioni sono dello stesso tipo, ovvero rappresentano lo stesso concetto azionale, mentre una è di un tipo diverso. Un concetto azionale è un'entità linguistico-cognitiva corrispondente a un pattern di modifiche del mondo compiute da un agente (umano, animale o macchina), ed è generalizzabile a vari oggetti (o azioni). Si tratta di una rappresentazione cognitiva di un evento o di un processo che coinvolge, prototipicamente, un agente (chi compie l'azione), un tema o paziente (sul quale si esercita l'azione) e, talvolta, uno strumento, un destinatario o una destinazione. Un concetto azionale è produttivo, ovvero può applicarsi a un'ampia varietà di oggetti e si presenta in contesti diversi. L'associazione tra concetto azionale e verbo che lo descrive non è un rapporto di tipo uno-a-uno. Infatti, un concetto azionale può essere realizzato linguisticamente con più verbi (ad es. 'spostare una scatola' e 'spingere una scatola') e, viceversa, un verbo può rappresentare più concetti azionali distinti (ad es. 'aprire una porta' vs. 'aprire una noce'). Nell'individuare un concetto azionale, è importante concentrare l'attenzione su quali cambiamenti vengono compiuti dall'azione rappresentata, non sul verbo. Tra le seguenti quattro frasi, individua la frase che descrive l'azione di tipo diverso dalle altre tre. Esiste solo una risposta esatta, rispondi utilizzando esclusivamente il numero di riferimento della frase e nient'altro.

Few-shot prompts

Few-shot prompts are created by appending 4 examples to the three variants of zero-shot prompts; additionally, a fourth prompt with only examples and no description is provided.

1) I ragazzi spingono i carrelli lungo il binario del treno
2) La donna con gli occhiali da sole spinge l'anziana signora sulla sedia a rotelle
3) L'uomo spinge una carriola nel cortile della fattoria mentre parla con la donna
4) Il veterinario spinge lo stantuffo della siringa
Intruso: 4

1) Il ragazzo si tira su in ginocchio
2) L'uomo si spinge sulle braccia per alzarsi in piedi
3) Il ragazzo ferito si spinge sui gomiti
4) L'operatore spinge in basso la leva dell'ascensore
Intruso: 4

1) La donna spinge l'uomo sul letto per farlo sdraiare
2) Il veterinario spinge lo stantuffo della siringa
3) L'uomo armato sposta il compagno dietro di lui
4) Il marinaio sposta i corpi galleggianti con le mani
Intruso: 2

1) La donna sposta leggermente la tenda di perline
2) La ragazza abbassa la mano del ragazzo con la pistola
3) La donna spinge in alto il pannello di vetro
4) La donna preme un pulsante del suo orologio
Intruso: 4

1) La donna preme sul posacenere al centro del tavolo
2) Il ragazzo spinge le scope nel ripostiglio
3) Il ragazzo spinge il pulsante di rilascio della cintura di sicurezza
4) L'uomo di scatto chiama l'ascensore
Intruso: 2

B. Complete results

Model	short zero	short few	medium zero	medium few	long zero	long few	none few
minerva-7b-instruct-v1.0	0.000	0.263	0.004	0.579	0.014	0.540	0.256
mistral-7b-instruct-v0.3	0.126	0.435	0.189	0.449	0.340	0.509	0.435
qwen2.5-7b-instruct	0.105	0.126	0.263	0.239	0.277	0.267	0.126
aya-expanse-8b	0.151	0.379	0.228	0.418	0.253	0.382	0.379
llama-3.1-8b-instruct	**0.604**	**0.726**	**0.926**	**0.761**	**0.933**	**0.705**	**0.726**
gemma-2-9b	0.298	0.089	0.319	0.133	0.456	0.109	0.102
velvet-14b	0.004	0.000	0.018	0.000	0.014	0.000	0.000
qwen-2.5-72b-instruct	0.158	0.495	0.449	0.512	0.519	0.579	0.491

Table 4
Results for pairs of sentences with same verbs (binary choice)

Model	short zero	short few	medium zero	medium few	long zero	long few	none few
minerva-7b-instruct-v1.0	**1.000**	0.912	0.993	0.604	0.144	0.670	0.912
mistral-7b-instruct-v0.3	0.986	0.863	0.954	0.825	0.853	0.796	0.863
qwen2.5-7b-instruct	0.972	0.975	0.905	0.965	0.912	0.944	0.975
aya-expanse-8b	0.965	0.891	0.947	0.877	0.926	0.898	0.891
llama-3.1-8b-instruct	0.716	0.653	0.221	0.572	0.193	0.688	0.653
gemma-2-9b	0.846	0.723	0.870	0.898	0.761	0.674	0.839
velvet-14b	**1.000**	**1.000**	**0.996**	**1.000**	**1.000**	**1.000**	**1.000**
qwen-2.5-72b-instruct	0.982	0.919	0.905	0.902	0.846	0.870	0.919

Table 5
Results for pairs of sentences with different verbs (binary choice)

Model	short zero	short few	medium zero	medium few	long zero	long few	none few
minerva-7b-instruct-v1.0	0.000	0.228	0.000	0.228	0.000	0.246	0.228
mistral-7b-instruct-v0.3	0.386	0.298	0.263	0.281	0.246	0.316	0.298
qwen2.5-7b-instruct	0.316	0.439	0.333	0.439	0.281	0.439	0.456
aya-expanse-8b	0.386	0.421	0.333	0.439	0.368	0.439	0.421
llama-3.1-8b-instruct	0.281	0.333	0.246	0.281	0.246	0.263	0.333
gemma-2-9b	0.263	0.368	0.140	0.281	0.228	0.316	0.368
velvet-14B	0.263	0.246	0.211	0.211	0.263	0.281	0.246
qwen-2.5-72b-instruct	**0.596**	**0.632**	**0.561**	**0.632**	**0.579**	**0.632**	**0.632**

Table 6
Accuracy values for quadruples where the intruder is expressed by the `same` verb

Model	short zero	short few	medium zero	medium few	long zero	long few	none few
minerva-7b-instruct-v1.0	0.000	0.184	0.000	0.184	0.000	0.184	0.184
mistral-7b-instruct-v0.3	0.237	0.474	0.263	0.447	0.211	0.447	0.474
qwen2.5-7b-instruct	0.684	0.737	0.711	0.737	0.605	0.737	0.737
aya-expanse-8b	0.500	0.395	0.421	0.500	0.447	0.553	0.395
llama-3.1-8b-instruct	0.447	0.684	0.526	0.632	0.553	0.632	0.684
gemma-2-9b	0.316	0.658	0.289	0.658	0.447	0.684	0.658
velvet-14b	0.605	0.289	0.632	0.316	0.658	0.184	0.289
qwen-2.5-72b-instruct	**0.868**	**0.868**	**0.842**	**0.895**	**0.868**	**0.895**	**0.895**

Table 7
Accuracy values for quadruples where the intruder is expressed by a `different` verb