MM-IGLU-IT: Multi-Modal Interactive Grounded Language Understanding in Italian

Federico Borazio(✉)[1 [0009−0000−0193−2131]], Claudiu Daniel Hromei(✉)[1 [0009−0000−8204−5023]], Elisa Passone[1 [0009−0009−0318−7441]], Danilo Croce(✉)[1 [0000−0001−9111−1950]], and Roberto Basili[1 [0000−0001−5140−0694]]

Department of Enterprise Engineering, University of Rome Tor Vergata, Italy
{borazio,hromei,passone}@ing.uniroma2.it
{croce,basili}@info.uniroma2.it

Abstract. This paper explores Interactive Grounded Language Understanding (IGLU) within Human-Robot Interaction (HRI). Here, a robot interprets user commands related to its environment, determining if a specific command can be executed. When ambiguities or incomplete data arise, the robot asks relevant clarification questions. Current models, trained on English datasets, leverage multi-modal and end-to-end capabilities by fine-tuning architectures like LLaVA. These models combine a Visual Encoder, processing images of the environment, with Large Language Models (LLMs) encoding user requests, enabling agents to discern command executability and seek clarifications when necessary. While many LLMs are inherently multi-lingual, fine-tuning them on English-only datasets limits their application in other languages, such as Italian. To address this, we developed MM-IGLU-IT, a dataset for Multi-Modal Interactive Grounded Language Understanding in Italian. This dataset was created by automatically translating existing large-scale datasets and manually validating them for accuracy, resulting in over 6,800 command examples. Training a model like LLaVA, fine-tuned over a multi-lingual base model such as LLaMA2, allowed us to achieve comparable performance in both English and Italian. This resource is released on a dedicated GitHub page at https://github.com/crux82/MM-IGLU-IT and we hope it will advance multi-modal models in the Italian language, providing a valuable resource for ongoing research.

Keywords: Human-Robot Interaction · Interactive Grounded Language Understanding · Large Language Models · Multi-Modality

1 Introduction

In recent years, significant progress has been made in developing models for text comprehension and interpretation. These models can answer questions, generate narratives, and interpret natural language and images [14, 23, 35, 39]. Additionally, there is a growing interest in models focused on interpreting commands, evidenced by the proliferation of Large Language Models (LLMs) like ChatGPT.

In robotics, while models excel at understanding human instructions, interpreting commands in real-world scenarios adds complexity. For instance, commands may be ambiguous, requiring clarification to ensure correct actions. The Interactive Grounded Language Understanding (IGLU) [13] task at NeurIPS 2022 showcased advances in natural language command interpretation. Here, "Understanding" involves interpreting a user's command, checking its feasibility, and generating an appropriate response. In this task, a human "Architect" gives commands to a robotic "Builder" in a Minecraft-like environment, such as "*Place 3 green blocks vertically above the red block*". The robot must determine if the commands are executable or need clarification. To address the challenges of command interpretation in Human-Robot Interaction (HRI), two primary strategies have emerged: *(i)* utilizing a Knowledge Base (KB) to store comprehensive entity information and integrate this knowledge into models; *(ii)* leveraging images to capture intricate details of nearby objects, including their spatial relationships, shapes, and colors, to develop end-to-end systems capable of accurately understanding and responding to user queries. These two strategies can be used independently or combined within a more complex system. Inspired by the latter method, the MM-IGLU [11] resource expands the original IGLU resource by incorporating images that depict block arrangements with their respective colors, paired with textual commands and expected responses from the robot, which include phrases like "*Yes, I can execute this command*" or "*Do you want me to move the red block positioned on the right or left*?". This enhancement enables the adoption of multi-modal approaches that merge visual perception encoding, obtained through advanced computer vision techniques, with text encoding. Notable examples of this class include ChatGPT4 [26], Flamingo [1], LLaVA [20], CogAgent [9] and Idefics [17].

However, MM-IGLU is exclusively for English language data, allowing for the training and evaluation of multi-modal LLMs only in English. To enable multi-modal LLMs to work with the Italian language, we created MM-IGLU-IT by translating and, most importantly, manually validating the available English data from MM-IGLU into Italian, resulting in more than 6,800 examples of commands. By training a multi-modal LLM, such as LLaVA, fine-tuned over a multi-lingual base model like LLaMA2 with the Italian data, we achieved comparable performance in both English and Italian. We hope this dataset will support the advancement of multi-modal models for the Italian language, providing a valuable resource for ongoing research in this domain.

In the rest of the paper, Section 2 describes the related work, Section 3 describes the resources and architectural frameworks used, Section 4 presents and discusses the experimental evaluation while Section 5 derives the conclusion.

2 Related Works

The generation of clarifying questions for human-robot interaction dates back to Winograd's foundational research [38]. Since then, many approaches have been developed, from human-made templates, such as cloze-type [8], rule-based

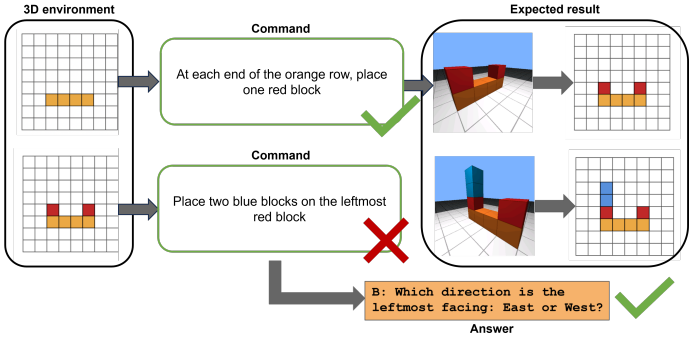

Fig. 1. Taken from the IGLU challenge description. *Top*: The architect's command was clear and no questions were needed, thus the Builder can execute it. *Bottom*: The word '*leftmost*' in the Command is ambiguous, so the Builder asks a clarifying question.

[24, 34], to semi-automatic questions [19, 33]. The most recent research has introduced Transformer-based techniques. This attention-based architecture, presented by [37], is an encoder-decoder architecture that has led to different model families. The encoder component, which may correspond to models like BERT [6], RoBERTa [21], and DeBERTa [7], encodes input sequences by using self-attention. In contrast, decoders, such as GPT [30], GPT-3 [4], and LLaMA [36], generate output sequences auto-regressively. LLaMA is a massive model with various applications to linguistic tasks, as shown in [10]. Examples of encoder-decoder models include T5 [31] and BART [18], which excel in tasks like translation, summarization, and question-answering. More recent research on generating clarifying questions has introduced transformer-based techniques. In [16], BERT is trained on an inverted SQuAD dataset [32], generating questions from provided text and answers. A different approach is expressed in [22] where GPT-2, used for the same dataset excluding answers, generates questions based purely on the context.

Despite these advancements, existing architectures focus on generating questions about a text but do not attempt to interact directly with the user to gather additional information in a specific environment. This limitation hinders the development of truly interactive systems that can dynamically engage with users. In the context of HRI, the successful interaction between the human and the robot is crucial. Effective collaboration requires clear roles and an understanding of each participant's position in the space [28]. For instance, [15] explores simulations of human behaviors where a robotic leader provides natural language commands, and the evaluation focuses on human task execution. However, this setup lacks full interactivity, as the human follower cannot ask questions but must follow the given instructions. Conversely, [25] investigates a dynamic interaction between a human and a robot capable of identifying when provided information is inadequate, a feature extended in [11].

The Interactive Grounded Language Understanding (IGLU) challenge, presented in [13], promotes research in Human-Robot Interaction, emphasizing collaboration via natural language. The challenge involves generating interactive agents that execute tasks using grounded language instructions in teamwork settings. Within IGLU, the "Architect" (Human Agent) instructs the "Builder" (AI Agent) on arranging colored blocks in a voxel environment. The Builder can seek clarifications if instructions are ambiguous, posing questions, as shown in Figure 1. In this context, interactions are single-turn: the Architect instructs, and the Builder acts or asks for clarity. More details about data gathering can be found in [2, 3, 12, 13]. However, the IGLU dataset lacks real-world images or natural language descriptions, and examples aren't categorized by command objectives, limiting the possibility of investigating multi-modal models. The work described in [11] aims to generate fully interactive systems based on Language Models, addressing these gaps by introducing multi-modal data and natural language descriptions. They integrate visual information, such as images of the environment, to explore unified visual and language systems. The paper tackles the task of Grounded Question Generation via a multi-modal approach integrating a Language Model based on LLaMA [36] with a Vision Model based on CLIP [29], merging both visual and textual data.

Although many models, including [36], are inherently multi-lingual, fine-tuning on English-only datasets limits their applicability to other languages. This work investigates the positive impact of creating an Italian dataset to enable effective evaluation and fine-tuning of multi-modal models in Italian. By developing MM-IGLU-IT, we aim to extend these models' capabilities to operate in Italian, thus broadening their applicability in human-robot interaction.

3 MM-IGLU-IT: An Italian Multi-Modal Dataset for Grounded Language Understanding

The original IGLU dataset primarily provided data that are numerical-only (such as the positions of the blocks and the numerical identifier of the color), which limited its direct applicability to multi-modal neural approaches that integrate vision and language, usually based on images of the environment. MM-IGLU [11] overcomes these limitations by incorporating images showing block configurations and natural language descriptions of the environment. As illustrated in Figure 2, an instruction like "*Break the green blocks*" cannot be executed if there are no green blocks in the environment, prompting the agent to seek clarification from the architect. The visual representation provided by these images enables the application of advanced computer vision techniques. Furthermore, MM-IGLU includes detailed textual descriptions of the blocks, such as "*There are no blue blocks, no yellow blocks, no green blocks, no orange blocks, eight purple blocks, four of which are on the ground, six red blocks, one of which is on the ground*". They are made by converting the three-dimensional block coordinates into detailed narratives that enumerate the number of blocks, classify them by color, and specify their positions, enhancing the model's ability to interpret and

respond to commands accurately. These linguistic descriptions allow the use of Large Language Models (LLMs) even in the absence of visual inputs, enhancing the dataset's versatility.

While multi-modal models like LLaVA can leverage LLMs agnostically, whether language-specific or multi-lingual, MM-IGLU's exclusive use of English data restricts the training and evaluation of models in other languages, such as Italian. To overcome this limitation, we developed MM-IGLU-IT by translating the English data from MM-IGLU into Italian and performing manual validation. This enabled the creation and assessment of multi-modal models in the Italian language. Inspired by the approach of [5], which translated the Visual Question Answering dataset into Italian, we utilized DeepL for the initial translation[1]. However, unlike [5], which validated only the test set, our process involved manual validation of the entire dataset—including training, validation, and test sets—by two annotators. This comprehensive validation ensured the integrity of the data, reducing the exposure of models to synthetic data and enhancing the overall quality of the training process.

Before completing validation, we evaluated the translation quality using the well-known BLEU scores [27] on the test set for both user commands and clarification questions. The BLEU-n scores for the commands are, in increasing lengths n of the target n-grams: 0.88, 0.83, 0.78 and 0.73, for n=1,2,3,4 respectively. For the questions, the corresponding scores are 0.95, 0.92, 0.48, and 0.39. High BLEU-1 scores suggest good overall translation quality, but the decline in scores with higher n-grams highlighted the need for further corrections. This prompted a manual validation of all translations to ensure accuracy. Consequently, MM-IGLU-IT encompasses over 6,800 examples, as detailed in Table 1. Each example includes an image depicting the arrangement and colors of blocks in the environment, accompanied by a command. If the command is not executable given the configuration, an expected clarification statement from the robot is included. Similar to IGLU and MM-IGLU, 13% of the examples require clarification. Furthermore, each Italian example is aligned with the original image and the corresponding command in English, supporting future cross-lingual research.

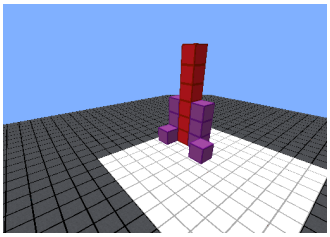

Fig. 2. An example of visual rendering of the environment, where the Instruction given by the Human is "*Break the green blocks*" and the expected answer is "*There are no green blocks, which blocks should I break?*".

The main sources of error that were corrected include incorrect verb conjugations (e.g., "*Rimuovete*" → "*Rimuovi*"), rephrasing expressions (e.g., "*Vista a Nord*" → "*Guardando a Nord*"), and fixing mistranslations (e.g., "*Towel*"

[1] Source accessed in March 2024 at https://www.deepl.com/it/translator.

Section	Instructions			Avg Len	
	#Exs	#Clear	#Amb	C	Q
Train	5,530	4,813	717	17.35	11.35
Val	615	531	84	16.39	10.79
Test	683	593	90	17.34	10.67

Table 1. Statistics of the datasets for total examples (#Exs), clear commands (#Clear), ambiguous commands (#Amb), and average word length for commands (C) and questions (Q).

→ "*Asciugamano*" instead of "*Torre*", i.e. "*Tower*"). For example, the English word "*block*" was often mistranslated as "*isolato*" (i.e. *city block*) instead of "*blocco*". Only about 1% of the commands were nonsensical, such as "*Facing north and green purple blocks will be destroyed*" which was translated to "*Rivolto a nord e blocchi verdi viola saranno distrutti*" despite not being actionable. We maintained ambiguous commands to highlight cases needing clarification questions and test the robustness of the neural models. For instance, "*Istruzione non chiara, cosa vuoi che faccia?*" translated from "*The command is not clear, what do you want me to do?*". It is interesting to note that verbs like *to destroy* were translated to Italian verbs *distruggere*, *cancellare*, or *rimuovere*, maintaining a broader linguistic variability than the original dataset. This variability can enhance an LLM's robustness by minimizing overfitting to specific verbs and actions. Similar to [11], for each command in the test set that exhibits ambiguity, we reassigned the same classification label of the original dataset specifying the type of information that the command lacks, prompting the need for a clarifying question. These categories include: BLOCK, indicating uncertainty about which block the command refers to, e.g., "*Which specific block do you mean?*", or COLOR, when clarification about the color of the block is required. We believe this categorization is very useful for understanding the need for additional information, but for more details, we refer to [11].

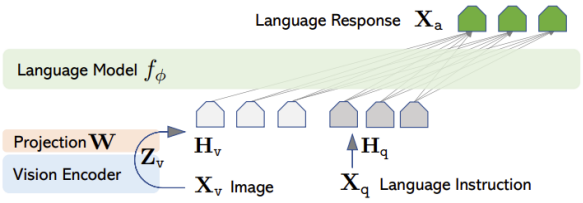

Fig. 3. The LLaVA network architecture, as presented in [20]

LLMs for Multi-modal IGLU. To address the above multi-modal task, we adopt the approach described in [11], utilizing the LLaVA framework [20]. This model integrates visual information from images of the environment using an image encoder like CLIP [29], with linguistic information using a Large Language

Model (LLM) such as LLaMA2 Chat-13b [36]. The model processes commands and generates textual outputs, such as "*Posso eseguirlo.*" (in English "*I can execute it*") if the command is executable, or a clarification request is needed, otherwise.

In practice, LLaVA combines visual models with linguistic models using a single-layer neural network called the Projector to align the visual model's output representation with the language model's input representation. The architecture, illustrated in Figure 3, shows X_v as the image and X_q as the input text. The Vision Encoder processes X_v to produce visual features Z_v, which are then projected by the Projection layer W to align with the language model's vector space, resulting in H_v. Simultaneously, the input text X_q is tokenized and embedded into the language model, producing H_q. These aligned visual and linguistic embeddings, H_v and H_q, are combined within the language model to generate a coherent language response X_a. This alignment is crucial for effective communication between the language and vision components of the model, enabling it to leverage both modalities effectively.

In this setup, the model is fine-tuned[2] by taking as input the tuple:

$$\langle \texttt{Introduction}, \texttt{Prompt}, \texttt{Image}, \texttt{Command} \rangle$$

The `Introduction` provides a contextual backdrop for the overarching task[3]:

> *In this virtual world reminiscent of Minecraft, you are a robotic entity equipped with the ability to move freely, and place or remove blocks within the environment. Imagine you are situated in the environment depicted in the image provided. Your task is to determine whether you can execute a given command based on the current configuration of the world. If you require additional information to carry out the command effectively, you should respond by asking relevant clarifying questions, such as inquiring about block colors, quantities, directions, or any other necessary details.*

The IGLU tasks can be modeled into two modalities: classification and generation. In the classification modality, the agent determines whether a command can be executed and responds with either "*Yes*" or "*No*". In the generation task, the agent generates a textual response to indicate whether the command can be executed and, if not, produces a clarification question. While these tasks only affect the `Prompt`, they could lead to two separate datasets. The `Prompt` element delineates the specific subtask at hand. For the classification task, it states:

> *Respond with 'Yes' if you can execute the command, or 'No' if additional information is required.*

For generation tasks, the prompt is:

[2] Initially, this model was tested in a zero-shot manner but it resulted in unstable outcomes, often leading to hallucinated answers. While most sentences generated were sensible, they typically failed to show an understanding of the need to perform actions within the environment, often miscounting blocks.

[3] All the following texts are translated in Italian when used in the model.

Answer with 'I can execute it' if the command is executable, or pose a pertinent clarifying question if further details are needed.

The `Image` token serves as a placeholder that the vision encoder subsequently replaces with X_v. Meanwhile, the `Command` represents the robotic directive. Thus, X_q is the concatenation of `Instruction`, `Prompt` and `Command`.

The model's output X_a conforms to a "*Yes/No*" structure for classification, or it produces the direct question for generation tasks or, again, the affirmative response "*I can execute it*". Inspired by the recent findings in [10, 11], which demonstrated the effective fusion of data from multiple tasks to guide the prompting of an LLM, we have introduced the capability for multi-modal models to train a single LLaVA model by combining data from both the classification and generation task prompts. This multi-task learning approach shows great potential, as we expect, based on findings from [10], that the tasks will complement and enhance each other's performance. In particular, the generation task might see improvements as the model implicitly specializes in the classification task. From a practical standpoint, it simply requires merging the training datasets generated from both modalities and ad hoc instructions.

In [11], the language model was based on LLaMA2 Chat-13b[4] with 13 billion parameters. In this work, we use the same LLaMA2 Chat-13b, since it has been partially trained on Italian data from previous versions, demonstrating high performance in processing Italian texts [10]. This choice ensures better adaptability and effectiveness for tasks in Italian, leveraging the strengths of both visual and linguistic modalities in a multi-modal framework.

4 Experimental Evaluation

In this section, we evaluate the performance of the proposed architecture in generating contextually grounded clarifications, providing insights into its understanding of instructions and its ability to identify missing information that can be transformed into queries. We utilize the LLaMA2-chat model as a generative decoder for the robot's responses, leveraging its multi-lingual capabilities to investigate and compare performance when fine-tuned on tasks using English or Italian data. Our analysis focuses on three main areas: *Quality of Generated Answers*, which assesses both the model's decision to refrain from asking questions and the nature of the questions it generates; *In-Depth Error Analysis*, which examines the model's limitations and areas of difficulty; and *End-to-End Question-Answer Generation*, which explores the capability of a holistic system to produce valid responses. Given the multi-lingual nature of the LLaMA model, we test different language combinations between English and Italian. Specifically, we compare the multi-modal model introduced in [11] and trained on the English dataset (`LLaMA2Chat-13b-EN`) with the model trained using the Italian dataset (`LLaMA2Chat-13b-IT`). This comparison allows us to assess the impact of language-specific training on model performance. As in [11], the linear

[4] https://huggingface.co/meta-llama/Llama-2-13b-chat-hf

projector, initially derived from LLaVA's release, was later completely re-tuned to achieve slight improvements in convergence. The hyper-parameters remain unchanged and are referenced in detail in that work.

Model name	Tr. Lan	Test Lan	F1 Pos	F1 Neg	M-F1
`LLaMA2Chat-13b-EN`	EN	EN	96.43%	67.16%	81.80%
`LLaMA2Chat-13b-EN`	EN	IT	70.07%	24.29%	47.18%
`LLaMA2Chat-13b-IT`	IT	IT	97.81%	66.67%	82.24%

Table 2. The Classification performance is divided into F1 of the positive class (the command is clear), F1 of the negative class (the command is ambiguous), and the Macro F1 of the two. The evaluation is divided into the Language of Training (Tr. Lan) and the Language of Testing (Test Lan).

Recognizing Ambiguous Commands. An interesting aspect of this evaluation is testing the multi-modal model introduced in [11], trained on the English dataset, on our Italian test set. This comparison helps us to understand whether translating the data into Italian and re-training the model is necessary or if the English-trained model, with the "emergent" capabilities of a multi-lingual LLM, is sufficient. First, we evaluate the system's ability to recognize ambiguous or problematic commands where clarifications are needed (without assessing the quality of these clarifications). We present results from various models applied to two tasks: *Classification* (determining whether to respond with "*Yes*" or "*No*") as shown in Table 2, and *Generation* (where the generated text is "*I can execute it*" or any clarification question) as shown in Table 3. Each model's language configuration, fine-tuned using the LLaVA framework, utilizes the same CLIP visual encoder, which remained "frozen" during the fine-tuning process. Being a binary classification task, we used Precision, Recall, and F1 metrics to evaluate the model's ability to say "*Yes*" or "*No*". The overall model performance is measured using Macro-F1 scores. Observing the results in Table 2, which includes model responses in the form of Yes/No, the `LLaMA2Chat-13b-EN` model reflects the results from [11] and is used as a reference. Its F1 Positive score is 96.43% when asserting its ability to recognize commands consistent with the environment. However, its performance decreases when ambiguous commands need clarifications, resulting in an overall Macro-F1 of 81.80%. This is plausible given the dataset's imbalance, with only 13% of cases expecting a "*No*". Leveraging the multi-lingual capabilities of the LLaMA2 model, we validated its behavior on Italian data without further tuning: performance dropped significantly (Macro-F1 of 47.18%), highlighting the difficulty in identifying ambiguous commands (F1 Neg is 24.29%). Nonetheless, this is an interesting result, as in 24.29% of cases, the system, likely inspired by some similarities between English and Italian terms like "*destroy*"/"*distruggere*" or "*blue*"/"*blu*" and the common vision model, manages to (rarely) respond correctly. Finally, a fairer evaluation was conducted by assessing the model trained on Italian-translated data, `LLaMA2Chat-13b-IT`, on the Italian test data: it is interesting to note that not

only this model does achieve comparable quality to the English version in correctly identifying when a command is executable or ambiguous, but it shows also a slightly higher Macro-F1. Moreover, the model's quality in correctly identifying an ambiguous command in the Italian test more than doubled compared to its English counterpart (F1 Neg: 66.67% vs 24.29%). Finally, we evaluated the models in a generation setup where the task is to produce a complete phrase, either affirming the command or generating a clarification if the command is ambiguous. Table 3 presents the results for this task. The `LLaMA2Chat-13b-EN` model, trained and tested in English, shows an F1 Positive score of 93.95% and an F1 Negative score of 47.89%, resulting in a Macro-F1 of 70.92%. This model serves as our baseline. When the same model is applied to the Italian test set without any further tuning, its performance drops significantly, with an F1 Positive score of 70.01%, an F1 Negative score of 0.00%, and a Macro-F1 of 35.00%. This indicates that the model struggles significantly with ambiguous commands in Italian, often failing to generate appropriate clarifications and instead defaulting to incorrect responses. In contrast, the `LLaMA2Chat-13b-IT` model, fine-tuned on the Italian dataset, performs comparably to the English baseline on the Italian test set. It achieves an F1 Positive score of 93.62% and an F1 Negative score of 44.16%, resulting in a Macro-F1 of 68.89%. In general, a model trained solely on Yes/No responses appears more effective in recognizing this specific task, suggesting, as in [11], that an effective system should still use a multi-step approach: first, determine if the command is ambiguous, and second, generate the necessary clarification if the command is recognized as ambiguous. In summary, the results indicate that while the LLaMA2Chat-13b model trained in English can somewhat handle Italian data due to its multi-lingual capabilities, its performance is mostly divergent and it is significantly enhanced when specifically fine-tuned on the Italian dataset. This underscores the importance of localized training for achieving high performance in different languages, demonstrating the necessity and effectiveness of our MM-IGLU-IT dataset.

Model name	**Tr. Lan**	**Test Lan**	**F1 Pos**	**F1 Neg**	**M-F1**
`LLaMA2Chat-13b-EN`	EN	EN	93.95%	47.89%	70.92%
`LLaMA2Chat-13b-EN`	EN	IT	70.01%	0.00%	35.00%
`LLaMA2Chat-13b-IT`	IT	IT	93.62%	44.16%	68.89%

Table 3. The Generation performance is divided into F1 of the positive class (the command is clear), F1 of the negative class (the command is ambiguous), and the Macro F1 of the two.

Evaluating the Generated Clarifications. To evaluate the quality of the generated clarifications, we used the same approach as in [11]. Instead of measuring the quality of generations in terms of exact accuracy or BLEU scores, we isolated 90 instances where requests were generated: we assessed them using the Relaxed Accuracy metric. This metric determines the percentage of cases where, despite deviations from the original, the generated questions effectively addressed the ambiguity. If the generated query resolved the ambiguity, it was deemed cor-

Category	LLaMA2Chat-13b-IT	LLaMA2Chat-13b-EN
Block	30.00%	60.00%
Vertical-Horizontal	50.00%	70.00%
Number	66.67%	55.56%
Square	62.63%	65.79%
Color	33.34%	66.67%
Direction	60.00%	80.00%
Block Missing	27.28%	54.55%
Complete	97.63%	97.11%
Overall	92.34%	93.24%

Table 4. The categories of "missing" information in the command identified in this work. Each category is described by a question example. A Relaxed Accuracy is computed for each category on the test set.

rect; otherwise, it was incorrect. Building on the categorizations introduced in the original MM-IGLU [11], we further analyzed the system's effectiveness in addressing specific missing information classes. In Table 4, Relaxed Accuracy values for `LLaMA2Chat-13b-IT` are reported along with those for the original English model (`LLaMA2Chat-13b-EN`), divided by the meta-categories of the questions introduced in the original paper [11]. The results indicate that in both cases, the system achieves a Relaxed Accuracy between 92.34% and 93.24%. These comparable results highlight the utility of MM-IGLU-IT, demonstrating that over 90% of the agent's requests help the hypothetical human user understand what information is missing.

Score	Utility	Fluency
1	*Incorrect classification*	*Not Italian or random Italian words*
2	*The clarification suggests awareness of the task but misses some key aspects*	*Italian with grammatical errors*
3	*Perfect*	*Perfect*

Table 5. Scores for the Utility and Fluency metrics from 1 to 3, where both need to be maximized.

To better understand the utility and naturalness of the generated clarification requests, we enlisted two (human) external annotators unfamiliar with the project's specifics. They received both system-generated and gold-standard examples requiring clarifications, without any indication of the source, shuffled in a random order. Each annotator rated the clarifications on two dimensions: *Utility* and *Fluency*. Utility was scored between 1 and 3 based on the guidelines in the second column of Table 5, capturing the effectiveness of the clarification. Fluency was scored between 1 and 3 based on the criteria in the third column of Table 5, assessing the quality of the Italian writing[5]. The results, presented

[5] The inter-annotator agreement was judged to be very good, with a Pearson correlation of 0.81 for Utility and 0.83 for Fluency.

in Table 6, show that the `LLaMA2chat-13b-IT` model achieved the highest Utility score of 2.79 (out of 3), reflecting its ability to generate relevant questions and address important missing information, albeit with occasional inaccuracies. In terms of Fluency scores, all models performed very well: 2.98 for the Gold Standard annotation and 2.99 for the `LLaMA2chat-13b-IT` model. The generated clarifications are straightforward enough to appear even more useful than those suggested by the original annotators. For example, for the command "*Distruggi 1 blocco e mettine altri 3 in fila*"[6], the expected output is simply "*Distruggere quale blocco?*"[7], while our `LLaMA2-chat-13b-IT` model produces a much more comprehensive question, addressing all crucial points (missing information): "*Quale specifico blocco devo distruggere e quale colore/posizione/direzione deve avere la fila di 3 blocchi?*"[8].

Dataset	**Language**	**Utility**	**Fluency**
Gold standard	EN	2.16	2.91
`LLaMA2Chat-13b-EN`	EN	2.73	2.99
Gold standard	IT	2.69	2.98
`LLaMA2chat-13b-IT`	IT	2.79	2.99

Table 6. Utility and Fluency results for the Gold Standard and the Multi-Modal model (LLaMA2chat-13b-IT).

5 Conclusions

In this paper, we addressed the complexities of Interactive Grounded Language Understanding (IGLU) within the scope of Human-Robot Interaction (HRI). Our investigation focused on the robot's ability to comprehend and execute user instructions, particularly in scenarios with ambiguities or incomplete information. Leveraging the existing MM-IGLU resource, which aims to bridge gaps between user intent and robot understanding, we expanded its applicability to the Italian language. This involved translating and manually validating both commands and clarification questions to ensure accuracy. Our contribution lies in adapting the MM-IGLU resource to Italian and demonstrating that pre-training on English data alone is insufficient for optimal performance. The study showed that fine-tuning the model on Italian commands significantly enhances its effectiveness, underscoring the necessity of language-specific training for multi-modal models. Future research should explore the transition from controlled, synthetic environments to more dynamic and realistic settings. While current computer vision techniques provide robust tools, real-world scenarios pose unique challenges that need addressing. Additionally, evaluating large-scale Multi-Modal LLMs, such as GPT-4, in zero-shot learning scenarios could yield valuable insights.

[6] In English "*Destroy 1 block and build another 3 in a row*"
[7] In English "*Destroy which one block?*"
[8] In English "*Which specific block should I destroy, and what color/direction/position should the three-block row be?*"

Acknowledgements

Claudiu Daniel Hromei is a Ph.D. student enrolled in the National Ph.D. in Artificial Intelligence, XXXVII cycle, course on *Health and life sciences*, organized by the Università Campus Bio-Medico di Roma. We acknowledge financial support from the PNRR MUR project PE0000013-FAIR.

References

1. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J., Borgeaud, S., Brock, A., Nematzadeh, A., Sharifzadeh, S., Binkowski, M., Barreira, R., Vinyals, O., Zisserman, A., Simonyan, K.: Flamingo: a visual language model for few-shot learning (2022), https://arxiv.org/abs/2204.14198
2. Aliannejadi, M., Kiseleva, J., Chuklin, A., Dalton, J., Burtsev, M.: Building and evaluating open-domain dialogue corpora with clarifying questions. In: Moens, M.F., Huang, X., Specia, L., Yih, S.W.t. (eds.) Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. pp. 4473–4484. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic (Nov 2021). https://doi.org/10.18653/v1/2021.emnlp-main.367
3. Aliannejadi, M., Zamani, H., Crestani, F., Croft, W.B.: Asking clarifying questions in open-domain information-seeking conversations. CoRR **abs/1907.06554** (2019), http://arxiv.org/abs/1907.06554
4. Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language models are few-shot learners. CoRR **abs/2005.14165** (2020)
5. Croce, D., Passaro, L.C., Lenci, A., Basili, R.: Gqa-it: Italian question answering on image scene graphs. In: Fersini, E., Passarotti, M., Patti, V. (eds.) Proceedings of the Eighth Italian Conference on Computational Linguistics, CLiC-it 2021, Milan, Italy, January 26-28, 2022. CEUR Workshop Proceedings, vol. 3033. CEUR-WS.org (2021), http://ceur-ws.org/Vol-3033/paper42.pdf
6. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Burstein, J., Doran, C., Solorio, T. (eds.) Proceedings of the NAACL 2019. pp. 4171–4186 (2019)
7. He, P., Liu, X., Gao, J., Chen, W.: Deberta: decoding-enhanced bert with disentangled attention. In: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021 (2021)
8. Hermann, K.M., Kocisky, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., Blunsom, P.: Teaching machines to read and comprehend. In: Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 28. Curran Associates, Inc. (2015)
9. Hong, W., Wang, W., Lv, Q., Xu, J., Yu, W., Ji, J., Wang, Y., Wang, Z., Zhang, Y., Li, J., Xu, B., Dong, Y., Ding, M., Tang, J.: Cogagent: A visual language model for gui agents (2023), https://arxiv.org/abs/2312.08914

10. Hromei, C.D., Croce, D., Basile, V., Basili, R.: ExtremITA at EVALITA 2023: Multi-Task Sustainable Scaling to Large Language Models at its Extreme. In: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023). CEUR.org, Parma, Italy (September 2023)
11. Hromei, C.D., Margiotta, D., Croce, D., Basili, R.: MM-IGLU: Multi-modal interactive grounded language understanding. In: Calzolari, N., Kan, M.Y., Hoste, V., Lenci, A., Sakti, S., Xue, N. (eds.) Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). pp. 11440–11451. ELRA and ICCL, Torino, Italia (May 2024), https://aclanthology.org/2024.lrec-main.1000
12. Kiseleva, J., Li, Z., Aliannejadi, M., Mohanty, S., ter Hoeve, M., Burtsev, M., Skrynnik, A., Zholus, A., Panov, A., Srinet, K., et al.: Interactive grounded language understanding in a collaborative environment: Iglu 2021. In: NeurIPS 2021 Competitions and Demonstrations Track. pp. 146–161. PMLR (2022)
13. Kiseleva, J., Skrynnik, A., Zholus, A., Mohanty, S., Arabzadeh, N., Côté, M.A., Aliannejadi, M., Teruel, M., Li, Z., Burtsev, M., ter Hoeve, M., Volovikova, Z., Panov, A., Sun, Y., Srinet, K., Szlam, A., Awadallah, A.: Iglu 2022: Interactive grounded language understanding in a collaborative environment at neurips 2022 (2022)
14. Koh, J.Y., Salakhutdinov, R., Fried, D.: Grounding language models to images for multimodal inputs and outputs. In: Proceedings of the 40th International Conference on Machine Learning. ICML'23, JMLR.org (2023)
15. Kojima, N., Suhr, A., Artzi, Y.: Continual learning for grounded instruction generation by observing human following behavior. Transactions of the Association for Computational Linguistics **9** (2021). https://doi.org/10.1162/tacl_a_00428, https://aclanthology.org/2021.tacl-1.77
16. Kriangchaivech, K., Wangperawong, A.: Question generation by transformers. CoRR **abs/1909.05017** (2019)
17. Laurençon, H., Saulnier, L., Tronchon, L., Bekman, S., Singh, A., Lozhkov, A., Wang, T., Karamcheti, S., Rush, A.M., Kiela, D., Cord, M., Sanh, V.: Obelics: An open web-scale filtered dataset of interleaved image-text documents (2023), https://arxiv.org/abs/2306.16527
18. Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L.: BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. CoRR **abs/1910.13461** (2019)
19. Liu, D., Lin, C.: Sherlock: a semi-automatic quiz generation system using linked data. In: International Semantic Web Conference (Posters & Demos), 9–12. Citeseer (2014)
20. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) Advances in Neural Information Processing Systems. vol. 36, pp. 34892–34916. Curran Associates, Inc. (2023)
21. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized BERT pretraining approach. CoRR **abs/1907.11692** (2019)
22. Lopez, L.E., Cruz, D.K., Cruz, J.C.B., Cheng, C.: Transformer-based end-to-end question generation. CoRR **abs/2005.01107** (2020)

23. Mirowski, P., Mathewson, K.W., Pittman, J., Evans, R.: Co-writing screenplays and theatre scripts with language models: An evaluation by industry professionals (2022), https://arxiv.org/abs/2209.14958
24. Mitkov, R., Ha, L.A.: Computer-aided generation of multiple-choice tests. In: Proceedings of the HLT-NAACL 03 Workshop on Building Educational Applications Using Natural Language Processing. pp. 17–22 (2003), https://aclanthology.org/W03-0203
25. Narayan-Chen, A., Graber, C., Das, M., Islam, M.R., Dan, S., Natarajan, S., Doppa, J.R., Hockenmaier, J., Palmer, M., Roth, D.: Towards problem solving agents that communicate and learn. In: Proceedings of the First Workshop on Language Grounding for Robotics. Association for Computational Linguistics, Vancouver, Canada (Aug 2017). https://doi.org/10.18653/v1/W17-2812, https://aclanthology.org/W17-2812
26. OpenAI: Gpt-4 technical report (2023)
27. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Isabelle, P., Charniak, E., Lin, D. (eds.) Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. pp. 311–318. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA (Jul 2002). https://doi.org/10.3115/1073083.1073135, https://aclanthology.org/P02-1040
28. Pustejovsky, J., Krishnaswamy, N.: Multimodal semantics for affordances and actions. In: Kurosu, M. (ed.) Human-Computer Interaction. Theoretical Approaches and Design Methods. pp. 137–160. Springer International Publishing, Cham (2022)
29. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. CoRR **abs/2103.00020** (2021), https://arxiv.org/abs/2103.00020
30. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al.: Improving language understanding by generative pre-training (2018)
31. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res. **21**, 140:1–140:67 (2020), http://jmlr.org/papers/v21/20-074.html
32. Rajpurkar, P., Zhang, J., Lopyrev, K., Liang, P.: SQuAD: 100,000+ questions for machine comprehension of text. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. pp. 2383–2392. Association for Computational Linguistics, Austin, Texas (Nov 2016). https://doi.org/10.18653/v1/D16-1264, https://aclanthology.org/D16-1264
33. Rey, G.A..C., Alexopoulos, I.., Damljanovic, P.., Damova, D.., Li, M.., ;, N., Devedzic, V.: Semi-automatic generation of quizzes and learning artifacts from linked data. In: Conference: Proceedings of the 2nd International Workshop on Learning and Education with the Web of Data (LiLe2012), co-located with the World Wide Web Conference (WWW2012) (2012)
34. Rus, V., Wyse, B., Piwek, P., Lintean, M., Stoyanchev, S., Moldovan, C.: The first question generation shared task evaluation challenge. In: Kelleher, J., Namee, B.M., Sluis, I.v.d. (eds.) Proceedings of the 6th International Natural Language Generation Conference. Association for Computational Linguistics (Jul 2010), https://aclanthology.org/W10-4234
35. Su, D., Xu, Y., Winata, G.I., Xu, P., Kim, H., Liu, Z., Fung, P.: Generalizing question answering system with pre-trained language model fine-tuning. In:

Proceedings of the 2nd Workshop on Machine Reading for Question Answering. Association for Computational Linguistics, Hong Kong, China (Nov 2019). https://doi.org/10.18653/v1/D19-5827, https://aclanthology.org/D19-5827
36. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., Lample, G.: Llama: Open and efficient foundation language models (2023), https://arxiv.org/abs/2302.13971
37. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017)
38. Winograd, T.: Procedures as a representation for data in a computer program for understanding natural language. Technical report, MASSACHUSETTS INST OF TECH CAMBRIDGE PROJECT MAC (1971)
39. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models (2023), https://arxiv.org/abs/2304.10592