MM-IGLU: Multi-Modal Interactive Grounded Language Understanding

Claudiu D. Hromei, Daniele Margiotta, Danilo Croce, Roberto Basili
Department of Enterprise Engineering, University of Rome Tor Vergata, Italy
hromei@ing.uniroma2.it,
daniele.margiotta@uniroma2.it,
{croce,basili}@info.uniroma2.it

Abstract

This paper explores Interactive Grounded Language Understanding (IGLU) challenges within Human-Robot Interaction (HRI). In this setting, a robot interprets user commands related to its environment, aiming to discern whether a specific command can be executed. If faced with ambiguities or incomplete data, the robot poses relevant clarification questions. Drawing from the NeurIPS 2022 IGLU competition, we enrich the dataset by introducing our multi-modal data and natural language descriptions in *MM-IGLU: Multi-Modal Interactive Grounded Language Understanding*. Utilizing a BART-based model that integrates the user's statement with the environment's description, and a cutting-edge Multi-Modal Large Language Model that merges both visual and textual data, we offer a valuable resource for ongoing research in the domain. Additionally, we discuss the evaluation methods for such tasks, highlighting potential limitations imposed by traditional string-match-based evaluations on this intricate multi-modal challenge. Moreover, we provide an evaluation benchmark based on human judgment to address the limits and capabilities of such baseline models. This resource is released on a dedicated GitHub repository at https://github.com/crux82/MM-IGLU.

Keywords: Human-Robot Interaction, Interactive Grounded Language Understanding, Large Language Models, Multi-Modality, Generative AI

1. Introduction

In recent years, there has been a surge in the development of models for text understanding and interpretation. Many of these models have been designed to answer questions, generate narratives, or facilitate natural language or image interpretations (Su et al., 2019; Mirowski et al., 2022; Koh et al., 2023; Zhu et al., 2023). Additionally, there has been a growing interest in models tailored for interpreting commands, as demonstrated by the proliferation of Large Language Models (LLMs) such as ChatGPT. In robotics, while models excel at understanding human instructions, real-world command interpretation introduces complexity. For instance, executing the command "*Take the book on the black chair*" requires the robot to discern what a "*chair*" is and which one is "*black*" among multiple options. Ambiguous commands further necessitate clarifications to ensure the correct action. Recently, the field of natural language command interpretation has seen a significant advancement, marked by the inception of the Interactive Grounded Language Understanding (IGLU) competition (Kiseleva et al., 2022b), held at NeurIPS 2022. "Understanding" in our context refers to the process of comprehending a user's command, assessing its feasibility in the given environment, and generating an appropriate response or request based on that assessment. In the IGLU competition, a human "Architect" issues natural language commands in English to a robotic "Builder", tasking it with determining whether the given commands are executable or necessitate further clarification through questions. The competition's simulated environment consists of a world composed of colored blocks, reminiscent of the popular game Minecraft, where commands like "*Place 3 green blocks vertically above the red block*" are issued, assuming, for instance, the existence of only one red block in the world. This kind of scenario is a typical example of Embodied Cognition (Anderson, 2003). In order to address these tasks, two primary strategies are evident: *i)* employing a Knowledge Base (KB) to archive detailed information about entities and subsequently infusing this knowledge into a model; or *ii)* leveraging images of the real world to grasp intricate details of nearby objects, encompassing their positions, shapes, and colors, among other characteristics. Notably, the second approach holds significant promise for the development of end-to-end systems, especially within the robotic domain. Such systems might perceive reality, albeit approximately, yet are proficient in interpreting user requests, discerning their feasibility, and responding appropriately.

In this paper, we journey into the development of models that seamlessly integrate visual perceptions of the environment with a natural language command in English. In a scenario like the IGLU competition, depending on its assessment, the model either responds with "*I can execute it.*" signifying the action's feasibility, or it generates a relevant inquiry

LREC-COLING 2024, pages 11440–11451
20-25 May, 2024.

for more clarity. Existing methodologies predominantly bifurcate into two primary classes. Firstly, those that practice "Textification", where they utilize natural language to articulate environmental perceptions (Hromei et al., 2022; Kiseleva et al., 2022b). These methods juxtapose linguistic representation with potential commands to utilize classifiers or Large Language Models (LLMs). Additionally, some systems implement a multi-modal approach that merges visual perception encoding, obtained through advanced computer vision techniques, such as Projections as in (Merullo et al., 2023), with text encoding. Notable examples of this class include ChatGPT4 (OpenAI, 2023), Flamingo (Alayrac et al., 2022), and LLaVA (Liu et al., 2023). These systems directly activate the LLMs, negating the need for distinct 'textification' modules.

In the context of the IGLU competition, where the provided data is predominantly text-based and relies on synthetic object descriptions delineated through matrices showcasing box colors and positions, our primary contribution lies in the expansion and enhancement of these resources. Our goal is to bridge the gap between the existing data and the needs of both textification-based and multi-modal grounded language understanding systems. To that end, we introduce *MM-IGLU: Multi-Modal Interactive Grounded Language Understanding*, an enriched version of the foundational IGLU dataset (Mohanty et al., 2022; Kiseleva et al., 2022a).

This refined dataset offers not merely text, but multi-modal information, encompassing both actual images depicting the world pertaining to each command and a meticulously constructed textified environment description. This approach ensures a controlled syntax, devoid of hallucinations, thereby granting a more robust dataset for a comprehensive array of methodologies.

Furthermore, we employed two distinct strong baselines: one leveraging the "textified image descriptions" and another that is purely multi-modal. In terms of model applications, this endeavor marks the inaugural application of leading models like LLaMA to the task, as well as the newly introduced multi-modal architectures, notably LLaVA (Liu et al., 2023). LLaVA, a large multi-modal model, adeptly unifies a vision encoder with an LLM, catering to a broad spectrum of visual and language understanding tasks. As such, our dataset is complemented by two robust baselines that serve as exemplars for this task. Finally, given the intricate nature of Grounded Language Understanding, we also introduce a more refined evaluation benchmark. This surpasses conventional metrics, extending beyond traditional standards like BLEU or Cider, ensuring a more holistic assessment of systems' capabilities.

In the rest, Section 2 provides an analysis of the literature, Section 3 presents the resource and the architectures proposed, Section 4 discusses the evaluation with an error analysis, while Section 5 derives some conclusions.

2. Related Work

The Transformer architecture, presented by Vaswani et al., 2017, divides into two main components, leading to different model families. The encoder, with models like BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and DeBERTa (He et al., 2021), encodes input sequences using self-attention. In contrast, decoders, such as GPT (Radford et al., 2018), GPT-3 (Brown et al., 2020), and LLaMA (Touvron et al., 2023), auto-regressively produce output sequences. LLaMA is a massive model that has been recently applied to diverse linguistic tasks, as shown in (Hromei et al., 2023).

Beyond these, Encoder-Decoder models like T5 (Raffel et al., 2020) and BART (Lewis et al., 2019) merge both components, excelling in tasks like translation, summarization, and question-answering. BART, especially, is trained to denoise corrupted text, enhancing its understanding and reasoning about text structure and content. In GrUT (Hromei et al., 2022), BART is trained to interpret robot commands in an automated house by grounding them using Frame Semantics (Fillmore, 1985). Given a command like "*Place the book on the black chair*" and an environmental description, BART produces a specific logical form, such as PLACING(THEME(B1), GOAL(C1)), where B1 denotes the book and C1 the black chair.

The generation of clarifying questions for human-robot interaction has deep roots, starting with Winograd's foundational research (Winograd, 1971). Numerous approaches have emerged, from human-made templates, such as cloze-type (Hermann et al., 2015), rule-based (Mitkov and Ha, 2003; Rus et al., 2010), to semi-automatic questions (Rey et al., 2012; Liu and Lin, 2014). Recent advancements introduced Transformer-based techniques, notably in (Kriangchaivech and Wangperawong, 2019), where BERT is trained on an inverted SQuAD dataset (Rajpurkar et al., 2016), generating questions from provided text and answers. Alternatively, (Lopez et al., 2020) uses GPT-2 for the same dataset, excluding answers, and producing questions based purely on the context.

All the aforementioned architectures focus on generating questions about a contextual text but do not attempt to interact with the user to gather additional information. In this paper, we aim to enrich the IGLU dataset with multi-modal information and to evaluate two different approaches for solving the task of Grounded Question Generation: *i)* a simple application of the BART architecture (Lewis et al., 2019) that relies solely on the command and

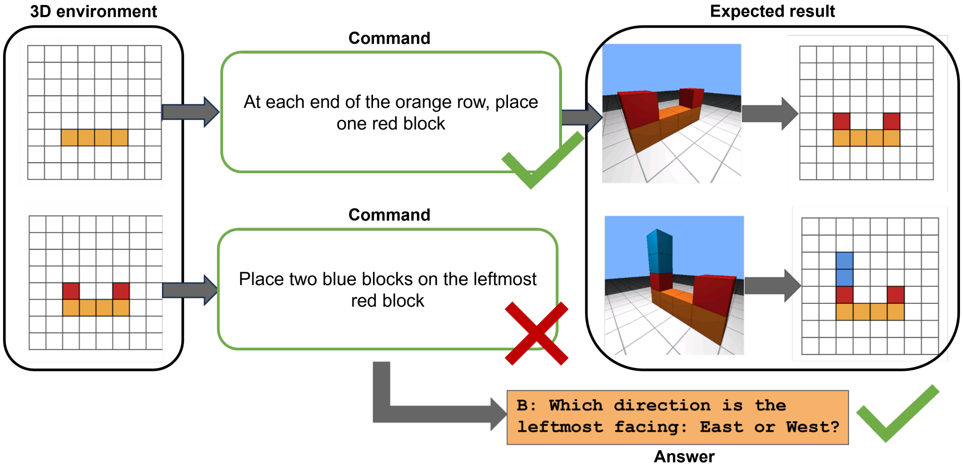
3D environment
Command
At each end of the orange row, place one red block
Expected result
Command
Place two blue blocks on the leftmost red block
B: Which direction is the leftmost facing: East or West?
Answer

Figure 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.

the natural language description of the environment (the textification enrichment we release); *ii)* a multi-modal model, integrating a Language Model based on LLaMA (Touvron et al., 2023) with a Vision Model based on CLIP (Radford et al., 2021), that relies on the images of the environment we release and the natural language command.

One key aspect in this scenario is the interaction between the Human and the Robot, as in the Human-Robot Interaction (HCI) field. For a perfect collaboration, the understanding of the roles of each interlocutor and their positioning in the space (Pustejovsky and Krishnaswamy, 2022) is crucial. In (Kojima et al., 2021), simulations of human behaviors are used to generate instructions in a collaborative setting involving a robotic leader and a human follower executing tasks in a specific environment. The robot provides natural language commands, and the evaluation focuses on human task execution. Notably, while the system generates the commands, the interaction is not fully interactive; the human follower cannot ask questions but must follow the given instructions. Conversely, Narayan-Chen et al., 2017 investigates a dynamic interaction between two agents: a human conveying information and a robot adapting to tasks and responding immediately. The robot in this study can identify when given information is inadequate, a feature we see in the IGLU dataset and plan to extend. In summary, the current landscape of human-agent interaction predominantly revolves around multi-turn or single-turn interactions in highly collaborative environments. In these settings, one agent issues commands in natural language, and the other agent executes them. Our contribution to this field is to incorporate natural language textual descriptions to enable fully interactive systems based on Language Models. Furthermore, we intend to integrate visual information, such as images of the environment, to explore unified Visual and Language systems, as elucidated in the literature (Abramson et al., 2020).

The IGLU competition. The IGLU challenge, presented in (Kiseleva et al., 2022b), promotes Human-Robot Interaction research, emphasizing collaboration via natural language. Its objective is crafting interactive agents adept at executing 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, while manipulating blocks, can seek clarifications if instructions are ambiguous. This challenge bridges Natural Language Understanding and Generation (NLU/G) and Reinforcement Learning (RL). This study concentrates on Grounded Question Generation. When the Builder receives commands, it determines the clarity of the information. If necessary, it poses questions, such as "*Which direction is the leftmost facing: east or west*?" to clarify ambiguities, as illustrated in Figure 1. The Builder's tasks encompass classification (deciding to ask) and ranking (choosing the best question from a predefined list). In IGLU's context, interactions are single-turn: the Architect instructs, and the Builder acts or asks for clarity. All required details are within this exchange, devoid of a broader narrative. Data was sourced from Amazon Mechanical Turk, where participants, placed in ongoing games, wrote commands based on specific 3D environments and objectives. The next agent decided on the adequacy of the previous command, leading to the collection of clarification questions. These questions were ranked by relevance. More details can be found in the collecting data papers (Aliannejadi et al., 2019, 2021; Kiseleva et al., 2022a,b).

While the IGLU dataset stands as a significant asset for the task, it solely provides command/question pairs coupled with an artificial representation of the environment, characterized by a three-dimensional matrix detailing block coordinates. It lacks real-world images or natural language descriptions vital for multi-modal model training. Furthermore, examples aren't categorized by command objectives, hindering a comprehensive model evaluation. To address these gaps, we introduce a resource delineated in the subsequent section.

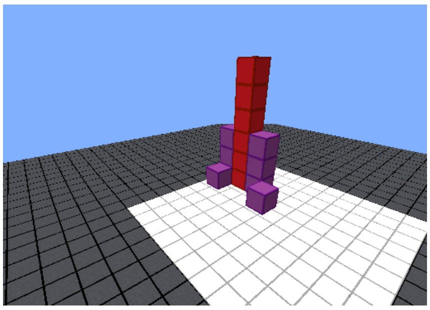

Figure 2: An example of visual rendering of the environment, where the Instruction given by the Human is "*Break the green blocks*" and the expected answer is "*There are no green blocks, which blocks should I break?*".

3. Multi-Modal Interactive Grounded Language Understanding

In this section, we introduce the Multi-Modal dataset and outline two paradigms to tackle the IGLU task: employing a linguistic-only model and deploying a unified Visual-Language Model, serving as strong baselines.

Building the MM-IGLU. To enrich the IGLU dataset with multi-modal evidence, we introduced two additional dimensions for each command-question pair: *i*) images depicting the block configurations in the environment, and *ii*) environment descriptions in natural language.

Images are generated by extending the `grid-world` tool made available in the competition to initialize an empty environment from the JSON-like description provided by IGLU and automatically place one block at a time in the correct position. We carefully selected a single viewpoint for the agent's perspective within the simulated world, positioning it at a slightly elevated angle from the ground to ensure the visibility of as many as possible blocks present, as illustrated in Figure 2. All the images have a resolution of $256 \text{x} 256$ pixels.

The description process converts the three-dimensional matrix detailing block coordinates into natural language descriptions, specifically English, using fixed templates. This textual representation enumerates the number of blocks present, classifying them by color and specifying the count of blocks resting on the ground. For instance, the textified description corresponding to Figure 2 is:

"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." (1)

This description is formulated in a synthetic language, devoid of hallucinations, and serves as a surrogate for visual input, elucidating the context in which the model operates.

Section	Instructions #Exs	#Clear	#Amb	Avg Len C	Q
Train	$5,530$	$4,813$	717	18.60	12.25
Val	615	531	84	17.42	11.46
Test	683	594	89	18.76	11.69

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").

We provide a different split of the overall dataset into Training, Validation and Testing sets (the latter not available during or after the competition), as reported in Table 1. Out of the $5,530$ training commands, 717 (12.97%) were annotated as "Ambiguous" (i.e., they require at least one question to advance in the task) while $4,813$ $(87,03\%)$ were considered "Clear" instructions (the task can be executed). The average length of clarifying questions is approximately 12 words, indicating that these questions tend to be quite specific. The same trend can be observed for the Validation set with 615 total commands and the Testing set with 683 total commands.

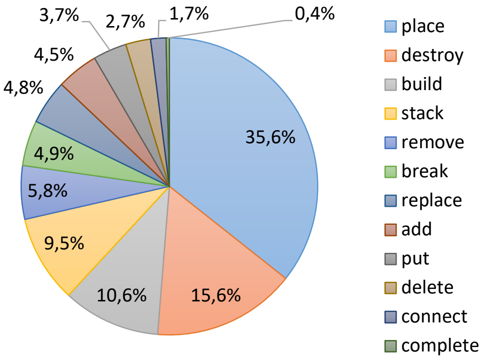

Figure 3: Command meta-categories in percentage.

Additionally, we introduce meta-categories to each question and command in the test set, to enable a more comprehensive analysis for the evaluation of answers generated by models, such as LLMs.

In our test set, commands can be categorized based on the actions they instruct the agent to perform, primarily determined by verbs relating to either placing or removing blocks in the environment. As depicted in Figure 3, the action *to place* is the most prevalent, constituting 36% of the commands, while *to destroy* accounts for 15%. A notable disparity exists among the actions, with commands for placing new blocks dominating at 66%, compared to the 34% that instruct block removal.

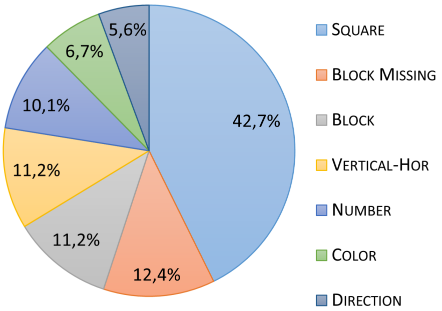

Figure 4: Question meta-categories in percentage.

For each command in the test set that exhibits ambiguity, we have appended a classification label specifying the type of information that the command lacks as in Figure 4, 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 ?*"; Vertical-Horizontal, when there's ambiguity in the block's vertical or horizontal alignment, e.g., "*How are they arranged? Vertical or horizontal?*"; Color, when clarity on the block's color is required, e.g.,"*Which color should the block be?*"; Direction, if the block's orientation is unclear, e.g.,"*In which direction? What is the orientation?*"; Block Missing, when the referenced blocks are absent in the environment, e.g.,"*There is no red block*"; Number, when it's uncertain how many blocks the command pertains to, e.g.,"*How many blocks? Or how long?*"; and Square, if it's ambiguous where the block should be placed, e.g.,"*Where should I place the blocks?*". For example, the instance in Figure 2 would be assigned to the Block Missing class as the command refers to a non-existing block.

LLMs for Multi-modal IGLU. We utilize advanced Transformer-based models and Large Language Models (LLMs), specifically tailored for sequence-to-sequence tasks. Whether processing an image or textual description, the model's objective remains consistent: taking an input (either text or text paired with an image) and producing a precise output (in natural language). Within the IGLU framework here considered, there's an integration of two tasks: *classification* and *generation*. The classification task prompts the model to decide whether the given command is executable based on the context provided. If affirmative, the model would confirm with a "*Yes*". Otherwise, it signals a "*No*". The generation task, on the other hand, steps in when the command lacks clarity. The model then formulates a pertinent question to gain the required clarity. There are two ways to approach these tasks during inference. The first is a *two-step method*: initially ascertain the executability of the command and, if deemed incomplete, subsequently generate the clarifying question. The alternative is a *monolithic* strategy: upon receiving the command, the model either assures with "*I can execute it*" or directly poses the needed question.

In particular, for evaluating systems that leverage world descriptions, we trained BART (Lewis et al., 2019), a state-of-the-art Transformer-based model, on natural language map descriptions. In our approach, we merge via concatenation the environment description (e.g., example 1) with natural language commands, forming the input for BART's interpretation. By incorporating detailed information like the color of ground blocks into this textual context, the model is better equipped to address ambiguities. For instance, when given a combined input of a description in example 1 and a command like "*Break the green blocks.*", BART can discern both the description and the command. If any clarity is needed, it generates pertinent questions, such as "*There are no green blocks, which blocks should I break?*". It's crucial to highlight that BART's ability to generate these context-aware questions stems from its comprehension of the description, which serves as a detailed description of the Minecraft-like environment and effectively acts as a substitute for visual cues. To train the model for the classification task, it is provided with the concatenated input of $\langle$`Description`, `Command`$\rangle$, and the expected output is "*Yes/No*". For the generation task, the same concatenated input is used, but the anticipated output is either a generated question or the affirmative response "*I can execute it*".

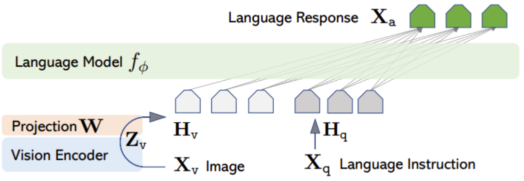

Figure 5: The LLaVA network architecture, taken from (Liu et al., 2023)

The above approach based purely on textual data is generalized by Multi-Modal LLMs that combine a LLM for understanding natural language commands and generating responses, along with a Vision Model for encoding and representing images. This integration allows us to bridge the gap between textual instructions and visual information, enabling the model to perform tasks based on both modalities.

We adopted the Multi-modal approach upon the Large Language and Vision Assistant framework (LLaVA) introduced by (Liu et al., 2023). LLaVA integrates foundation visual models with linguistic models, employing a single-layer neural network, known as Projector, to align the output representation from the visual model with the input representation from the language model. The architecture of LLaVA is depicted in Figure 5. In this figure, X_v and X_q represent the image and input text, respectively, while H_v and H_q denote their embedding representations, which have already been aligned with the language model. The input text X_q undergoes tokenization, while the image X_v passes through the Vision Encoder and the Projection layer W to ensure alignment with the Language Model vector space. This alignment is crucial for effective communication between the language and vision components of the model, enabling it to leverage both modalities.

In this setup, the model is fine-tuned[1] (and later used) 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. It reads: "*In this virtual world reminiscent of Minecraft, you are a robotic entity equipped with the ability to move freely, 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 `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: "*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 tasks, or it produces the direct question for generation tasks or, again, the affirmative response "*I can execute it*". Lastly, inspired by the recent findings in (Hromei et al., 2023), 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 is promising, as we anticipate, based on (Hromei et al., 2023), that the tasks will mutually benefit each other. 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.

4. Experimental Evaluation

In this section, we evaluate the ability of the two proposed baselines to generate contextually grounded clarifications, offering insights into their comprehension of instructions and the identification of missing information that can be turned into a query. Our analysis will focus on three key areas: *Quality of Generated Answers*, assessing both the model's decision to refrain from asking questions and the nature of the questions they produce; *In-Depth Error Analysis*, providing a comprehensive examination of model limitations and understanding areas of difficulty; and *End-to-End Question-Answer Generation*, exploring the capability of a holistic system to produce valid responses. It's important to emphasize that variations in BLUE scores between two sentences, A and A', don't necessarily reflect differences in relevance to our task; as such, a manual evaluation of generated questions is undertaken.

Experimental Setup. The BART-IGLU linguistic-only model utilizes the BART-base version from Huggingface. It is trained with concatenated environment descriptions and user utterances as inputs. We employed the standard Cross-Entropy loss for text generation without any specific adaptations for classification tasks. In the LLaVA architecture, two essential components are to be selected: the LLM and the Vision Module for encoding images. After some initial experimentation, we leaned towards the combination of CLIP (Radford et al., 2021) for the Vision Module and LLaMA2 (Touvron et al., 2023) for the LLM, as they yielded superior results on the development set. While there were other contenders like Vicuna (Zheng et al., 2023) and Guanaco (Dettmers et al., 2023), their results are not detailed here due to space constraints. The LLaMA2, built using the HuggingFace framework, comes with size variations ($7b$ vs $13b$) and can be the base or chat version. Meanwhile, the Vision Module is solidly anchored in CLIP and remained unchanged throughout our experiments[2]. The pro-

[1] 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.

[2] We adhered to the fine-tuning process described in

jector, a single-layer Feedforward Neural Network, was initially based on LLaVA's release but later re-tuned from scratch due to slightly improved convergence.

The hyper-parameters are optimized on the development set and summarized in Table 7 at the end of the paper.

Model name	Type	Tr.task	F1 Pos	F1 Neg	M-F1
BART-IGLU	TO	Gen	93.76%	54.55%	74.36%
LLaMA2-7b	MM	Class	96.04%	62.05%	79.05%
LLaMA2Chat-7b	MM	Class	96.42%	67.16%	81.79%
LLaMA2-13b	MM	Class	96.19%	64.12%	80.16%
LLaMA2Chat-13b	MM	Class	**96.43%**	**67.16%**	**81.80%**
LLaMA2Chat-13b	MM	MT	96.35%	66.17%	81.26%

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 Type TO stands for Textual-Only and MM stands for Multi-Modal. MT here stands for Multi-Task training, i.e. the union of the Classification dataset and the Generation one, using ad hoc instructions.

Evaluating the Question Generation Process. We detail the results from various Language Models applied to Classification (deciding whether to ask or not) and Generation (determining what to ask) tasks. All models fine-tuned with the LLAVA framework employ the same CLIP visual encoder, which remained "frozen" during fine-tuning. We evaluated models based on Macro-F1 scores, considering the phrase "*I can execute it.*" as the positive response to the command "Can you execute this command?"

For Classification (as seen in Table 2), the Text-Only BART-IGLU, reliant on command and environment description, scored a Macro F1 of 74.36%, hindered mainly when determining to pose questions (negative F1). This model, originally trained to generate questions, showed this classification as a side effect of its training. We also fine-tuned BART for classification and multitask modes, yielding similar results to the generation mode. Notably, the model equates "*Yes*" with "*I can execute it.*", showing limited multitasking generalization. A direct comparison with the system participating in the IGLU competition is unfeasible due to unavailable test data. The top three competition systems scored F1 scores of 76.6%, 76.1%, and 75.4%. Our local test result, despite differing from the online test set, is comparable to top-performing systems, underscoring our model's end-to-end nature.

With Multi-Modal (MM) solutions, using LLaMA2 checkpoints from the LLaVA framework, we initially observed subpar results, potentially due to LLaVA's training with real images rather than Minecraft-style environments. Thus, we focused on Meta-LLaMA2 models, fine-tuned for command Classification. These exhibited improved performance, particularly the Chat variants (which were previously fine-tuned on instruction), with the $7b$ version rising from 79.05% to 81.79% and the $13b$ from 80.16% to 81.80%. Notably, no marked difference existed between the two sizes. We further trained the LLaMA2Chat-13b version using Multi-Task training, alternating Classification, and Generation tasks. This showed a marginal decline in Classification performance, reaching 81.26% Macro-F1 (but later a better performance in generation).

For the generation task, in contrast to the original competition, where the task was framed as retrieving a possible question from a set pool, our approach here poses a greater challenge. We treat it as a genuine generation task, where a question is deemed correct only if it matches the dataset's expected query. While most commands are straightforward, leading to higher positive F1 scores with responses like "*I can execute it.*", the real challenge arises with ambiguous commands. Simpler models, like BART, struggle here, exhibiting much lower performance than LLMs. As shown in Table 3, the approach using BART-IGLU achieves a Macro-F1 of 50.45%. However, the Multi-Modal approach, integrating visual cues, surpasses it by nearly 20%. Notably, BART's performance significantly lags behind LLaMAs, especially in ambiguous contexts. To gauge the Unified MM-model system's efficacy, we evaluated its generated text quality. Recognizing that even a single word variance from the gold standard could deem a text inaccurate under our stringent initial metric, we turned to standard sequence evaluation metrics, like BLEU. The results show declining scores as the n-gram count increases, with BLEU1 at 0.124, BLEU2 = 0.063, BLEU3 = 0.0373, and BLEU4 dropping to a mere 0.026. However, this quantitative measure, originally designed for evaluating tasks like Machine Translation, can be overly restrictive. For instance, in response to a command such as "*Destroy all the red blocks*", a system that answers, "*The map contains no red blocks*" may share no common terms with a response like "*I don't see any elements of the requested color*", reaching a BLEU score of 0.

Model name	Type	Tr. task	F1 Pos	F1 Neg	M-F1
BART-IGLU	TO	Gen	93.76%	7.14%	50.45%
LLaMA2Chat-13b	MM	Gen	93.90%	45.26%	69.58%
LLaMA2Chat-13b	MM	MT	**93.95%**	**47.89%**	**70.92%**

Table 3: 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 Type TO stands for Textual-Only and MM stands for Multi-Modal. MT here stands for Multi-Task training, i.e. the union of the Classification dataset and the Generation one, using ad hoc instructions.

(Liu et al., 2023) Their rationale was that CLIP performed well on their images without further tuning. While our images are in a "Minecraft-style", zero-shot descriptions requested from the model were accurate, suggesting no further fine-tuning of CLIP was necessary.

Category	BART-IGLU	MM-model
Block	38.46%	**60.00%**
Vertical-Horizontal	50.00%	**70.00%**
Number	**57.14%**	55.56%
Square	**77.14%**	65.79%
Color	50.00%	**66.67%**
Direction	22.23%	**80.00%**
Block Missing	**58.34%**	54.55%
Complete	**97.81%**	97.11%
Overall	92.54%	**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. The MM-model is based on LLAVA using LLaMA2Chat-13b.

Evaluating the generated commands. As a result, we conducted a qualitative assessment requiring manual analysis on a test set of 683 examples. From these, we isolated 89 instances where requests were generated. Introducing the Relaxed-Accuracy metric, we determined the percentage of cases where, despite deviations from the original, the generated questions effectively addressed ambiguity. If the generated query resolved the ambiguity, it was deemed correct, otherwise incorrect. Moreover, building on the categorizations introduced in Section 3 (Table 4), we further analyzed the system's effectiveness in addressing specific missing information classes. In Table 4, Relaxed Accuracy values for BART-IGLU and the Multi-Modal LLaMA model (based on LLaMA2Chat-13b) are presented. The Multi-Modal solution slightly excels in the Overall metric. Notably, for categories like Block, Vertical-Horizontal, Color, and Direction, MM-model shows superior performance, underscoring its capability to produce pertinent questions to clarify ambiguities. For example, for the command "*Destroy 1 block and build another 3 in a row*", the expected output is simply "*Destroy which one block?*" while the Multi-Modal model produces a much more comprehensive question, addressing all the crucial points (missing information): "*Which specific block should I destroy, and what color/direction/position should the three-block row be?*".

Conversely, BART-IGLU outperforms the MM-model in certain scenarios, possibly because the image presented during generation offers only a singular perspective, which might obscure some blocks or cells. BART-IGLU has access to a comprehensive (textual) description of the world. For the MM-model, the lowest result is obtained in the Block Missing category. In contrast, for the Direction category, the result increases from 22.23% to 80%, revealing greater robustness in understanding when information about the direction is needed. An interesting example is the command "*Destroy the yellow block. In its place put a 4-block vertical red column*" associated with a world containing only one yellow block. While the Gold Standard annotates the command as incomplete coupled with the question "*In which direction do we place the 4 red blocks in a vertical column?*", BART-IGLU generates a similar question but the MM-model can understand that there is no need to ask any questions because "*vertical column*" is not ambiguous.

Score	Utility	Fluency
1	*Incorrect classification*	*Not English or random English words*
2	*The clarification suggests awareness of the task but misses some key aspects*	*English 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.

In an additional experimental analysis, we addressed an inherent shortcoming in the Relaxed Accuracy metric. Even though it's relevant to the task, the annotators are aware of which clarifications come from the gold standard (and which are generated by the system). This could introduce biases, especially if they notice the system frequently producing accurate clarifications. To counteract this, we enlisted two external annotators unfamiliar with the project's specifics. They were given both the system-generated examples and the gold standard examples requiring clarifications, without any indication of the source. This approach was designed to minimize any bias. Each annotator was then instructed to rate the clarifications on two dimensions. First, they gauged *Utility*, assigning a score between 1 and 3 based on the guidelines in the second column of Table 5. This measure was meant to capture the effectiveness of the clarification concerning the task in a nuanced manner. Second, they evaluated *Fluency*, providing a score between 1 and 3 based on the criteria outlined in the third column of Table 5. This assessed the quality of the English writing, taking into account both grammatical and syntactical facets. The results are presented in Table 6, where the MM-model achieves the best Utility score of 2.73 (over 3), reflecting its ability to generate more relevant questions and address important missing information, though it is not without occasional inaccuracies. In terms of Fluency scores, all models perform very well: 2.91 for the Gold Standard annotation, 2.98 for the BART-IGLU model, and 2.99 for the MM-model. The Pearson's correlation between the two annotators is 0.64 for the Utility score and 0.81 for the Fluency. The results suggest that the model generates simple, effective and linguistically correct sentences (leveraging the power of the LLM) and is straightforward enough to seemingly be even more useful than the clarifications suggested by the original annotators.

Dataset	Utility	Fluency
Gold standard	2.16	2.91
BART-IGLU	2.37	2.98
MM-model	2.73	2.99

Table 6: Utility and Fluency results for the Gold Standard (GS), the linguistic-only BART-IGLU, and the Multi-Modal model (MM-model).

5. Conclusions

In this paper, we addressed the complexities of Interactive Grounded Language Understanding (IGLU) in the context of Human-Robot Interaction (HRI). Central to our investigation was the robot's capability to comprehend and act on user instructions, particularly when faced with ambiguities or incomplete information. Our response to these challenges was the development of clarification questions, aiming to resolve discrepancies between user intent and robot comprehension. Leveraging insights from the NeurIPS 2022 IGLU competition, we presented a dataset, fortified with our multi-modal data and natural language descriptions, that serves the research community as a tool for further exploration. The integration of a BART-based model with the Multi-Modal Large Language Model demonstrates the synergy between visual and textual data in the realm of IGLU.

Future directions should consider the transition from controlled, synthetic environments to more dynamic and realistic settings. Though computer vision provides robust tools, real-world scenarios introduce unique challenges. Additionally, evaluating large-scale Multi-Modal LLMs, such as GPT-4, in zero-shot learning scenarios can offer intriguing insights. Extending our methodology to accommodate advanced multi-modal data, like videos, is a prospective next step.

Code, Data and Links

The software for generating the images for the IGLU dataset is taken from https://github.com/iglu-contest/gridworld and adapted in order to position the viewpoint in a specific position. Moreover, a mapping of the color IDs from the IGLU dataset to the "gridworld" environment was necessary. Finally, the original IGLU training dataset can be downloaded from https://github.com/microsoft/iglu-datasets with an MIT license.

The models utilized in this paper can be downloaded from Huggingface:

- **BART-base**: https://huggingface.co/facebook/bart-base
- **LLaVA LLaMA2 13b**: https://huggingface.co/liuhaotian/llava-pretrain-llama-2-13b-chat
- **LLaMA2 13b**: https://huggingface.co/meta-llama/LLaMA-2-13b-hf
- **LLaMA2 13b Chat**: https://huggingface.co/meta-llama/Llama-2-13b-chat-hf
- **CLIP**: https://huggingface.co/openai/clip-vit-large-patch14

Parameter Name	BART-IGLU Value	LLaMA Value
Optimizer	AdamW	AdamW
Early_stopping_delta	$1 \cdot 10^{-3}$	None
Early_stopping_metric	eval_loss	None
Batch_size	16	16
Early_stopping_patience	2	None
Scheduler	Linear with warmup	Linear with warmup
Warmup Ratio	0.1	0.1
Max_length	128	2048
Learning rate	$3 \cdot 10^{-5}$	$2 \cdot 10^{-4}$
Epochs	50 (max)	10
Model Size	base	$7b$ & $13b$

Table 7: Summarization of the hyper-parameters for the BART and LLaMA Language Models.

Ethics Statements and Limitations

Training a model like LLaMA2 incurs significant computational costs, demanding hundreds of hours on a GPU. While we've implemented optimizations, such as applying the LoRA (Hu et al., 2021) technique with the Peft (Mangrulkar et al., 2022) package and mixed precision approximations, to expedite the process, training on a 16GB or 20GB GPU still necessitates substantial computational resources. This is further pronounced by the model's sentence processing time, averaging one second per sentence, which is relatively lengthy. In terms of the model's application, its heavy reliance on an LLM raises concerns about potential hallucination, where it might generate non-existent sentences or fragments. However, during inference, we've observed that it has consistently stayed within the boundaries of the Minecraft-like world. Nevertheless, a more comprehensive review is essential to validate this observation. To ensure the evaluation process remains untainted by external factors, additional experiments may be required. The dataset should not be part of the pre-training phase, as it is not publicly available on the web and must be downloaded from the competition page. A notable limitation of our model is its reliance on English-only fine-tuning datasets for commands and generated questions. This restricts its ability to handle languages other than English. Additionally, the model's synthetic images are derived from a game-like, simulated environment. Evaluating its performance with different languages and diverse environments would provide valuable insights.

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.

Bibliographical References

Josh Abramson, Arun Ahuja, Arthur Brussee, Federico Carnevale, Mary Cassin, Stephen Clark, Andrew Dudzik, Petko Georgiev, Aurelia Guy, Tim Harley, Felix Hill, Alden Hung, Zachary Kenton, Jessica Landon, Timothy P. Lillicrap, Kory W. Mathewson, Alistair Muldal, Adam Santoro, Nikolay Savinov, Vikrant Varma, Greg Wayne, Nathaniel Wong, Chen Yan, and Rui Zhu. 2020. Imitating interactive intelligence. *CoRR*, abs/2012.05672.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. 2022. Flamingo: a visual language model for few-shot learning.

Mohammad Aliannejadi, Julia Kiseleva, Aleksandr Chuklin, Jeff Dalton, and Mikhail Burtsev. 2021. Building and evaluating open-domain dialogue corpora with clarifying questions. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4473–4484, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mohammad Aliannejadi, Hamed Zamani, Fabio Crestani, and W. Bruce Croft. 2019. Asking clarifying questions in open-domain information-seeking conversations. *CoRR*, abs/1907.06554.

Michael L. Anderson. 2003. Embodied cognition: A field guide. *Artificial Intelligence*, 149(1):91–130.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *CoRR*, abs/2005.14165.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the NAACL 2019*, pages 4171–4186.

Charles J. Fillmore. 1985. Frames and the semantics of understanding. *Quaderni di Semantica*, 6(2):222–254.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: decoding-enhanced bert with disentangled attention. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Claudiu Daniel Hromei, Danilo Croce, Valerio Basile, and Roberto Basili. 2023. 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)*, Parma, Italy. CEUR.org.

Claudiu Daniel Hromei, Danilo Croce, and Roberto Basili. 2022. Grounding end-to-end architectures for semantic role labeling in human robot interaction. In *Proceedings of the Sixth Workshop on Natural Language for Artificial Intelligence (NL4AI 2022) co-located with 21th International Conference of the Italian Association for Artificial Intelligence (AI*IA 2022), Udine, November 30th, 2022*, volume 3287 of *CEUR Workshop Proceedings*, pages 24–38. CEUR-WS.org.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *CoRR*, abs/2106.09685.

Julia Kiseleva, Ziming Li, Mohammad Aliannejadi, Shrestha Mohanty, Maartje ter Hoeve, Mikhail Burtsev, Alexey Skrynnik, Artem Zholus, Aleksandr Panov, Kavya Srinet, et al. 2022a. Interactive grounded language understanding in a collaborative environment: Iglu 2021. In *NeurIPS 2021 Competitions and Demonstrations Track*, pages 146–161. PMLR.

Julia Kiseleva, Alexey Skrynnik, Artem Zholus, Shrestha Mohanty, Negar Arabzadeh, Marc-Alexandre Côté, Mohammad Aliannejadi, Milagro Teruel, Ziming Li, Mikhail Burtsev, Maartje ter Hoeve, Zoya Volovikova, Aleksandr Panov, Yuxuan Sun, Kavya Srinet, Arthur Szlam, and Ahmed Awadallah. 2022b. Iglu 2022: Interactive grounded language understanding in a collaborative environment at neurips 2022.

Jing Yu Koh, Ruslan Salakhutdinov, and Daniel Fried. 2023. Grounding language models to images for multimodal inputs and outputs.

Noriyuki Kojima, Alane Suhr, and Yoav Artzi. 2021. Continual learning for grounded instruction generation by observing human following behavior. *Transactions of the Association for Computational Linguistics*, 9.

Kettip Kriangchaivech and Artit Wangperawong. 2019. Question generation by transformers. *CoRR*, abs/1909.05017.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *CoRR*, abs/1910.13461.

D. Liu and C. Lin. 2014. Sherlock: a semi-automatic quiz generation system using linked data. In *International Semantic Web Conference (Posters & Demos), 9–12. Citeseer*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Luis Enrico Lopez, Diane Kathryn Cruz, Jan Christian Blaise Cruz, and Charibeth Cheng. 2020. Transformer-based end-to-end question generation. *CoRR*, abs/2005.01107.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, and Sayak Paul. 2022. Peft: State-of-the-art parameter-efficient finetuning methods. `https://github.com/huggingface/peft`.

Jack Merullo, Louis Castricato, Carsten Eickhoff, and Ellie Pavlick. 2023. Linearly mapping from image to text space.

Piotr Mirowski, Kory W. Mathewson, Jaylen Pittman, and Richard Evans. 2022. Co-writing screenplays and theatre scripts with language models: An evaluation by industry professionals.

Ruslan Mitkov and Le An Ha. 2003. Computer-aided generation of multiple-choice tests. In *Proceedings of the HLT-NAACL 03 Workshop on Building Educational Applications Using Natural Language Processing*, pages 17–22.

Shrestha Mohanty, Negar Arabzadeh, Milagro Teruel, Yuxuan Sun, Artem Zholus, Alexey Skrynnik, Mikhail Burtsev, Kavya Srinet, Aleksandr Panov, Arthur Szlam, et al. 2022. Collecting interactive multi-modal datasets for grounded language understanding. *arXiv preprint arXiv:2211.06552*.

Anjali Narayan-Chen, Colin Graber, Mayukh Das, Md Rakibul Islam, Soham Dan, Sriraam Natarajan, Janardhan Rao Doppa, Julia Hockenmaier, Martha Palmer, and Dan Roth. 2017. Towards problem solving agents that communicate and learn. In *Proceedings of the First Workshop on Language Grounding for Robotics*, Vancouver, Canada. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report.

Gyu-Min Park, Seong-Eun Hong, and Seong-Bae Park. 2022. Post-training with interrogative sentences for enhancing BART-based Korean question generator. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*.

James Pustejovsky and Nikhil Krishnaswamy. 2022. Multimodal semantics for affordances and actions. In *Human-Computer Interaction. Theoretical Approaches and Design Methods*, pages 137–160, Cham. Springer International Publishing.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

G. A. ; Celino Rey, I. ; Alexopoulos, P. ; Damljanovic, D. ; Damova, M. ; Li, N. ;, and V. Devedzic. 2012. 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)*.

Vasile Rus, Brendan Wyse, Paul Piwek, Mihai Lintean, Svetlana Stoyanchev, and Christian Moldovan. 2010. The first question generation shared task evaluation challenge. In *Proceedings of the 6th International Natural Language Generation Conference*. Association for Computational Linguistics.

Dan Su, Yan Xu, Genta Indra Winata, Peng Xu, Hyeondey Kim, Zihan Liu, and Pascale Fung. 2019. Generalizing question answering system with pre-trained language model fine-tuning. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, Hong Kong, China. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

Terry Winograd. 1971. Procedures as a representation for data in a computer program for understanding natural language. Technical report, MASSACHUSETTS INST OF TECH CAMBRIDGE PROJECT MAC.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models.