Evaluating Embeddable Language Models in Verbalizing Rule-based Inferences through Justifications

Bastien Dussard[1], Aurélie Clodic[1] and Guillaume Sarthou[1]

***Abstract*— While Language Models have shown promising performance, they still struggle with limitations regarding reasoning and are very token-sensitive. In contrast, knowledge-based systems, such as ontologies, allow for provable logically valid reasoning and provide explicit justifications regarding newly inferred knowledge. However, those justifications can be hard to understand for non-expert users given their formal syntax and their length. We investigated if language models could be considered as reliable tools for verbalizing such explanations, thus increasing explainability over reasoning output. This paper presents a reference evaluation of a set of embeddable language models on a task of translation from ontology formatted inferences and justifications into natural language sentences. We show that the order of justifications significantly decreases performance, whereas adding the inference rule as additional context significantly improves performance, leading to more reliable results.**

I. INTRODUCTION

Language Models (LMs) have shown promising results in a number of complex tasks in robotics [1], from task planning [2] to HRI conversational skills [3], enhancing interaction quality by having systems that answer in a more natural manner and can justify their own decisions. The best performing models for such tasks are Large Language Models (LLMs) which, due to their substantial size, are not easily embeddable on robotic platforms' GPU. Instead, they rely on cloud connectivity to query the LLM, which does not align with the "fully autonomous" view of robotic agents as they depend on access to an external tool [4]. However, techniques such as *"knowledge distillation"* allow to transfer the knowledge of large models into smaller ones, leading to the emergence of Small Language Models (SLMs) which are embeddable locally on robotic platforms.

Nevertheless, LMs have exhibited limitations regarding reasoning, both in natural language or formal logics premises [5]. These limitations can be critical in robotic contexts as erroneous reasoning may lead to poor decision-making, resulting in suboptimal or potentially harmful behavior. For decades, knowledge-based systems have been reliable tools for reasoning purposes [6], allowing to represent and to reason about explicit knowledge, ensuring logical validity and tractability. Among these approaches, logic-based knowledge reasoning provides explicit justifications for inference results but may fall short on transferring this expert knowledge to non-expert users. Indeed, such knowledge can be highly structured and formal, or domain-specific assuming prior-knowledge, leading to hindered understanding.

Thanks to their understanding of common knowledge and their capability to interpret structured data, LMs are interesting candidates for such a translation task, from expert structured knowledge to natural language. Indeed, thanks to a reliable reasoning process performed ahead, the LM is only prompted with relevant and logically valid information. Consequently, the model is only tasked to contextualize the inference with trustworthy justifications and verbalize it into a natural language paragraph, mitigating their sensitivity to irrelevant information.

However, LMs have been shown to be very sensitive both to token variations [7], [8] and the number of steps in the reasoning process [9]. In robotic applications, where reasoning conclusions should be translated to a human partner to enable collaborative decision making, one can wonder how the generated explanations would be impacted by such factors in terms of logical validity and completion of the conveyed information.

In this paper, we choose to focus on ontologies as the knowledge-based system, and Semantic Web Rule Language (SWRL) rules as the reasoning process to be translated. Unlike other knowledge-based systems that rely heavily on propositional logic and constraints, ontologies rely on more expressive logic (predicate/description) that allows for more complex and human-interpretable knowledge representation. Indeed, where propositional logic deals with statements that are either true or false, ontologies account for structured relationships and hierarchies between entities as well as quantification, and other advantages. Furthermore, they allow for reasoning to extract new knowledge automatically trough inference rules, providing explainability through justifications.

In this work, we investigate the use of LMs as tools for verbalization of rule-based inferences according to their justifications. Although the evaluated models have been selected to be embeddable on robotic platforms, the conclusions derived from this study could reasonably be applied to larger models. Given the aforementioned limitations of LMs regarding reasoning tasks, we believe that factors such as the number of premises [9], their order [10] and the LM's sensitivity to token variations [7], [8], would also influence performance in the evaluated task. Consequently, we posit the hypothesis that first, the number of axioms

*This work was supported by the ELSA (ANR-21-CE33-0019) and the HumFleet (ANR-23-CE33-0003) projects.

[1]LAAS-CNRS, Université de Toulouse, CNRS, Toulouse, France firstname.surname@laas.fr

in the justifications would impact the LM's performance. Second, we posit that the logical flow of the justifications (their order), would also impact performance. Lastly, since some works have shown that adding explicitly the rule in natural language increases performance [11], we posit that adding the SWRL rule as additional context could guide the LM's understanding and improve performance.

The main contribution of this paper is an **evaluation of embeddable Language Models** on a translation task, from ontology-formatted inferences and justifications to natural language sentences. The models are evaluated on a dataset containing rule-based inferences with their corresponding justifications and the SWRL rule. The aforementioned hypotheses are verified through two metrics assessing the validity and the completion of the explanations.

II. BACKGROUND

In this section, we provide a brief overview of ontologies and the main inference mechanisms used in this paper. **Ontologies** are semantic knowledge bases, allowing a formal and explicit specification of shared meaning of any concept. They can be used to represent both conceptual and grounded knowledge related to a given situation. Ontologies rely on three main semantic types: classes (concepts), individuals (instances), and properties (predicates). An ontology can be viewed as a semantic network where Resource Description Framework (RDF) Triples, the atomic structures in the RDF data model, follow a subject–predicate–object format to create branches in the semantic network (e.g. $robot\_1|Type|Robot$ or $robot\_1|isNextTo|table\_2$).

More than a pure vocabulary specification, ontologies can be enriched with **axioms** which can be used for both validation and inference mechanisms. Inferences are ways of deducing new knowledge from existing one and axioms. Common axioms include inverse, transitive, or functional axioms for properties, as well as equivalence or inheritance for classes. When an inference is made in an ontology, reasoners can provide justifications as subsets of semantic facts that are sufficient to support this result.

SWRL rules (Semantic Web Rule Language) were introduced to extend the set of Web Ontology Language (OWL) axioms to include Horn-like rules. The latter are of the form of an antecedent and a consequent, both being a set of atoms. If the antecedents hold, then the consequents hold too, leading to inferences such as to deduce that a robot can perform an action. Fig. 1 presents an example of such a rule, that states that the triple $a|canGrasp|o$ would be inferred if: a is an agent having a grasping capability, o is an object having a graspable disposition, that a can reach o, and that a has an end-effector component with an opening width greater than the object's holding-part's width. The justifications for a rule-inferred entailment correspond to semantic facts validating each atom in the rule.

III. RELATED WORK

Ontology Verbalization refers to the transformation of structured, formal representation of ontological knowledge into natural language text, either for domain experts to access knowledge more efficiently or for non-experts to be able to understand it (e.g. axiom $Mug|SubClassOf|Object$ can be translated to *"A mug is a type of object"*).

Over the years, several works have been proposed for such a task, mostly in the form of Controlled Natural Languages (CNL). CNL define a restricted and unambiguous subset of natural language (NL), while providing a correspondence with Description Logic (DL) syntax (ACE [12], SOS [13]). Among these works, the verbalizer proposed in ACE [14] allows for bi-directional translations through rules and a grammar, and supports all OWL axioms as well as partially SWRL. The output of such methods is a collection of unordered translated sentences, which are closer to the English language but only accounts for a single CNL sentence per axiom. However, some OWL axioms can be related to the same entities, and thus their CNL translations could be merged for more compact textual representations. For this purpose, some works proposed to apply Natural Language Generation techniques such as template and sentence aggregation to provide more linguistically fluent paragraphs (SWAT [15], NaturalOWL [16]). Despite the improved readability with paragraphs instead of mere sentences, those paragraphs may suffer from redundancies which can further impact understandability of the translated knowledge. In order to overcome this problem, "Semantic-level refinement" [17] was introduced to remove redundancies according to a set of rules, improving the degree of understanding of the generated paragraphs. However, this approach requires heavy processing and doesn't support all OWL axioms. In sum, all the previous works mostly rely on a constrained grammar and rigid rules, which limit the expressivity and thus the potential understandability of such translations.

With the rise of LLMs and their emergent capabilities, some works investigated the use of such models to perform those translations, leveraging their ability to produce more fluent text and their understanding of formal logics syntax. In [18], the authors investigated such a verbalization task with axiom translation between DL to NL. They submitted queries to the LLM with each axiom in the ontologies, along with their associated types (e.g. *FunctionalObjectProperty*). Their results demonstrate that the model captured the essence of DL axioms with a majority of partially accurate answers, but struggles with complex axioms. Where this work tackled a single axiom at a time, the work presented in [19] relied on CNL methods to verbalize concepts into a set of sentences, which were then provided to the fine-tuned LLM to paraphrase them into a paragraph. Both works exhibit promising results but they evaluated well-known ontologies, which may have been a part of the training dataset, indicating potential

Agent(?a), hasCapability(?a, ?c), GraspingCapability(?c), Object(?o), hasDisposition(?o, ?d), GraspableDisposition(?d), isReachableBy(?o,?a), EndEffector(?g), hasComponent(?a,?g), hasOpeningWidth(?g,?w1), hasHoldingPartWidth(?o,?w2), greaterThan(?w1,?w2) -> canGrasp(a, ?o)

Fig. 1: Example of SWRL rule to infer a *canGrasp* relation.

data contamination. Additionally, they rely on guiding the LM with either additional information (axiom type), or pre-computed CNL-based sentences. Thus, they do not investigate the LLM's ability to verbalize such knowledge in a raw manner and to reason over semantic links between axioms.

Inferences in ontologies come with justifications to provide explainability regarding deductions. However, those justifications can be rather long and are in structured format hardly understandable by non-expert user. Justification Explanation refers to the transformation of such logical inference traces into NL text, that accounts for the logical relations between verbalized axioms. Some works proposed to translate the justifications into NL and shorten them with rule-based proof trees [20], [21]. Such approaches tend to identify causal links between axioms to extract subsets easily translatable and re-concatenated using patterns to linguistically connect the clauses (e.g. "since", "thus", etc). However such translations lack of naturalness because of the rule-based re-concatenation.

While current works mostly rely on pre-processing and/or LMs fine-tuning, to the best of our knowledge, such a translation task has never been evaluated in a baseline case. Such an evaluation would allow to assess LMs' initial performance on a reference case, as well as highlighting the factors influencing it. Understanding the effects of those factors could then lead to more informed strategies for improving performance thanks to other tools.

IV. METHODS

In this section, we present the dataset generation process, the design of the evaluation conditions, as well as the methods used to measure performances. Our code and dataset are available online [1].

A. Dataset Generation

The task evaluated in this paper consists of translating semantic facts (justifications) that enabled the deduction of new knowledge (inference) into a natural language paragraph (explanation). Since we aim to establish a reference evaluation to assess the models' ability to translate new knowledge, generating a dataset allows to do so while targeting factors to investigate. This study tackles rule-based inferences, thus we designed four SWRL rules which have as antecedents the enabling conditions for a robot to be able to perform an action on an object, and as consequents the corresponding semantic fact which links the robot to the object (*canPush, canGrasp, canPerceive, canLift*). The antecedents consists of atoms related to the robot and its capability, the object and its disposition, a spatial property (e.g. proximity between the robot and the object), and a physical property (e.g. the gripper's opening width), as illustrated in Fig. 1. Several modalities will be evaluated in this work, to provide a more in depth understanding of factors influencing performance, and are presented below.

Complexity: Since for rule-based inferences the length of the justifications could vary, depending on the number of semantic triples which allow each atom to hold, one could wonder how the models would perform on the same inferences to explain, but with varying justifications lengths. Therefore, we introduced three complexity levels represented by increased lengths for the justifications of some atoms (e.g. the grasping capability comes from having a gripper (**easy**), plus a motion planning algorithm (**medium**), plus the gripper not holding anything (**hard**)). The complexity levels require respectively 10, 14, and 17 semantic triples to be retrieved. At this stage we have $4 \times 3 = 12$ inferences to be translated.

Variations: Given the sensitivity of language models to token variations, we introduced variations in the inference/justifications pairs. First, we introduced variations in some of the involved concepts by selecting common proper names that match the atom's semantic meanings in the rule (e.g. objects with a handle could be a *Mug, Suitcase, etc*). Second, all involved individuals have been anonymised to avoid any semantic cues by picking random numbers of 1 to 5 characters (e.g. *mug_1* would become *plw*). Generated identifiers have then been manually assessed to not have any meaning (e.g. identifier *eye* has been transformed into *yye*). Finally, any numerical values have been randomly picked while ensuring that they still verified the built-in atoms conditions. Twenty variations have been generated per inferences to be translated giving $4 \times 3 \times 20 = 240$ inferences.

Conditions: Since the inference of a fact results from the resolution of a SWRL rule in the form of justifications, the semantic information behind such an entailment is represented by the names of the classes/properties in the triples. However, the structure of the inference is not explicit, meaning that while the triples do show semantic information, some further reasoning is required to link variables together. We consider as our **baseline** condition the triples in the order they have been used to make the inference. This order directly matches the one of the used rule and should thus have a logical/intuitive flow. However, since SWRL reasoners can use different exploration methods to compute the justifications given their implementation, the order of the semantic triples could vary. Therefore, we decided to include a **shuffle** condition that randomly shuffles the order of the justifications. Finally, as the justification is the list of facts used to resolve a given rule, we introduced the **rule** condition that provides, in addition to the justifications, the used SWRL rule. By adding those three conditions, we have $4 \times 3 \times 20 \times 3 = 720$ inferences to translate.

Queries: Queries have been created using the task prompt provided in Fig. 2, a common Chain-of-Thought (CoT) prompting with 4-shots, making sure that none of the concepts used in the examples are a part of the evaluated questions, and the generated inference/justifications pair to be translated by the models. Each queried pair to translate was provided with the exact same four CoT examples, with one of them displayed on the figure (in red). A queried inference/justification pair on a *canGrasp* question (on baseline condition and easy complexity) is shown in blue.

[1] https://github.com/RIS-WITH/inference_explanation_benchmark

Task Prompt

You are an AI model specialized in explaining formal logical inferences and their justifications in natural language.
Task Overview: Convert inferences and justifications from SWRL rules into concise, clear, and natural language explanations. SWRL rules infer new facts in ontologies based on known facts and logical conditions.
Input: Inference: A statement derived from SWRL rules. Justifications: Triples specifying semantic relationships between entities or concepts. Rule: The SWRL rule used to make such an inference.
Output Requirements: Do not include any explanation break down or apologies in your response. Do not add any other text other than the generated explanations. Clarity: Write a single, fluent sentence or a short paragraph without any breakdown or bullet points. Completeness: Include all justifications into the explanation. Concept-Driven: Don't refer to individual names (e.g., indiv_1|Type|Human becomes 'The human').
Edge Case: For inferences involving chains, equivalent concepts, subclasses or subproperties, explicitly reflect these relationships in the explanation.

Chain-of-Thought (×4-shots)

Inference: pw|canSpeakWith|lnj
Justifications: pw|Type|Robot, pw|hasCapability|j, j|Type|VerbalCommunicationCapability, lnj|Type|Human, lnj|hasDisposition|zl, zl|Type|VerbalCommunicationDisposition, (isInFrontOf o hasAgentSeatedOn)|SubPropertyOf|isFacing, pw|isInFrontOf|hx, hx|hasAgentSeatedOn|lnj, lnj|isAvailable|boolean#true.
Rule: Robot(?r), hasCapability(?r,?c), VerbalCommunicationCapability(?c), Human(?a), hasDisposition(?a, ?d), VerbalCommunicationDisposition(?d), isFacing(?r,?a), isAvailable(?a, boolean#true) → canSpeakWith(?r, ?a)
Output: Let's translate. Translation is the robot can speak with the human because it has a verbal communication capability and the human, which is available, has a disposition for verbal communication. Additionally, the robot is in front of an object that has the human seated on, ensuring that it is facing it.

Inference to translate

Inference: v|canGrasp|yo
Justifications: v|Type|Pepper, Pepper|SubClassOf|Agent, v|hasCapability|onq, GraspingCapability|EquivalentTo|(isCapabilityOf some (hasComponent some Gripper)), onq|isCapabilityOf|v, v|hasComponent|gjqb, gjqb|Type|TwoFingerClaw, TwoFingerClaw|SubClassOf|Gripper, yo|Type|Shovel, Shovel|SubClassOf|Object, yo|hasDisposition|za, GraspableDisposition|EquivalentTo|(isDispositionOf some (hasPart some GraspablePart)), za|isDispositionOf|yo, yo|hasPart|h, h|Type|Grip, Grip|SubClassOf|GraspablePart, yo|isReachableBy|v, gjqb|Type|Gripper, Gripper|SubClassOf|EndEffector, v|hasComponent|gjqb, gjqb|hasOpeningWidth|integer#71, yo|hasHoldingPartWidth|integer#50, greaterThan(integer#71,integer#50).
Output: Let's translate. Translation is

Fig. 2: Task prompt (green), one of the four examples provided in Chain-of-Thought (red), and the queried inference to translate on easy complexity and baseline condition (blue).

B. Performance Metrics and Annotation Guidelines

The annotation process was conducted by a single expert annotator to ensure consistency between evaluations, providing a fair comparison between models. Two metrics were computed on each example: the correctness and the completeness.

Correctness: boolean value which represents if the generated explanation matches the inference to-be translated, meaning it does not contradict the semantic meaning of the justifications for each atom in the rule.

Completeness: percentage of concepts which were translated from the semantic triples into natural language in the generated explanation.

A simplified example of what an annotation looks like is shown in Fig. 3, with its corresponding evaluation. Since the evaluation metrics could be rather subjective, we decided on a set of guidelines, helping to reduce any bias and creating a more objective/consistent evaluation.

Correctness Guidelines:

- The generated explanation must not contain any individual names, rather only refer to their classes (e.g. *"the robot 'zsq' ..."* is considered incorrect).
- The only inferred relation should be the consequent of the rule, meaning the action possibility. In such a way, if the resulting sentence tends to introduce causality between other atoms, the latter is considered incorrect (e.g. *"the robot has the grasping capability because it can reach the object"*).
- The generated sentence should refer to the grounded situation and should not present a general possibility of action for the involved agent (e.g. *"The robot can grasp because ..."* is considered incorrect while *"The robot can grasp the mug because..."* is correct).
- The explanation must not contradict any of the atoms used in the inference (e.g. *"The robot can't grasp ..."*).
- The explanation is still considered correct if it does not mention every expected concept.

Completeness Guidelines:

- The completeness metric takes into account every occurrence of the expected concept, even if it is wrongly used in the sentence (e.g. *"the gripper can grasp"*, would be counted as one occurrence of the gripper concept).
- Since explanations target a conversational context, the concept could be present implicitly and still be counted, even if it is not referred to by its class (*"the object"*), but not in a general case(*"an object"*).

C. Experimental Setup

Models: Six models from three families have been selected. The selection criteria has been to only select models that can be run offline and not exceeding 14GB. Selected models [2] were llama3.2:3b, llama3.1:8b, gemma2:2b, gemma2:9b, mistral-nemo:12b, mistral-small:22b.

Queries: Ollama has been used to interact with the selected models locally. Since a short paragraph has been

[2]deepseek-r1:7b did not exist at the time of this study

Example queried inference/justification pair with generated explanation

Inference: zsq|canGrasp|qmef
Justifications: zsq|Type|Robot, sc|isCapabilityOf|zsq, zsq|hasComponent|dzl, dzl|Type|Gripper, qmef|isReachableBy|zsq, GraspingCapability|EquivalentTo|(isCapabilityOf some (hasComponent some Gripper))
Rule: Robot(?r), isCapabilityOf(?c,?r), GraspingCapability(?c), isReachableBy(?o,?r) → canGrasp(?r,?o)

Generated explanation: "The robot 'zsq' can grasp because it has a grasping capability, based on its gripper and being able to reach the object".

Evaluation of the generated translation

Correctness: False
Completeness: 'can grasp', 'grasping', 'gripper', 'is reachable' → 4/4

Fig. 3: Example of a simplified inference to translate with the inferred semantic triple, the justifications satisfying the rule's constraints and the SWRL rule (in blue). In the middle, an example generated explanation with inconsistencies (underlined), and below the corresponding evaluation (in yellow).

instructed to the models, the answers have been truncated at the first newline character.

Results analysis: A three-way analysis of variance (ANOVA) was performed to determine if complexity levels, conditions and models had a significant effect on both investigated metrics. The associated estimates and p-values are reported in the following sections.

V. RESULTS

The following section presents the results obtained through the evaluation according to the investigated factors. First, the performance on baseline condition and average complexity is studied. Second, the complexity factor is investigated, still on the baseline condition. Third, the baseline condition is compared to the shuffle condition. Last, the rule condition is compared to the baseline condition. The full results are provided in Appendix I.

A. How do the models perform on average complexity? (Baseline condition)

In order to have a first idea of performance to expect, average complexity is studied by not distinguishing between complexity levels. Fig. 4 presents the results for each model, on average question complexity on the baseline condition. Models are organized by families in pairs (in row), with the smaller versions on the left. Each histogram bin corresponds to the number of occurrences a model exhibited a given completeness score. The dashed lines correspond to the correctness score, meaning the ratio of correctly translated explanations.

Completeness: For inferences on a given rule, all variations correspond to semantically similar justifications, meaning that the concepts used are inter-changeable and thus should not impact the results if the models have a fine understanding of the concepts. However, as mentioned earlier, LMs tend to be sensitive to token variations, with such a conclusion observable in Fig. 4. Indeed, all models show different results given the selected concept sets, leading to distributions instead of a single value (histogram instead of a single bin). From this figure, we can also observe that the completeness distributions of larger models are more centred to the right (closer to 100% score), both among pairs and across all models. This result is supported by the completeness values shown in Appendix. I, with larger versions having higher mean and lower standard deviation values than their smaller versions. However, the improvement gap between versions is not equivalent across all families, with the gemma2:9b model performing substantially better than its smaller version. Additionally, even if the largest model of all (mistral-small:22b) seems to have the narrowest distribution, its smaller version performs similarly on the mean value, with half the size.

Correctness: Each pair of model shows a better correctness score with the larger version, indicated by the dashed line being more centred to the right (100% correctness) and as reported in Appendix. I. This result also supports the hypothesis that the size of the model impacts performance,

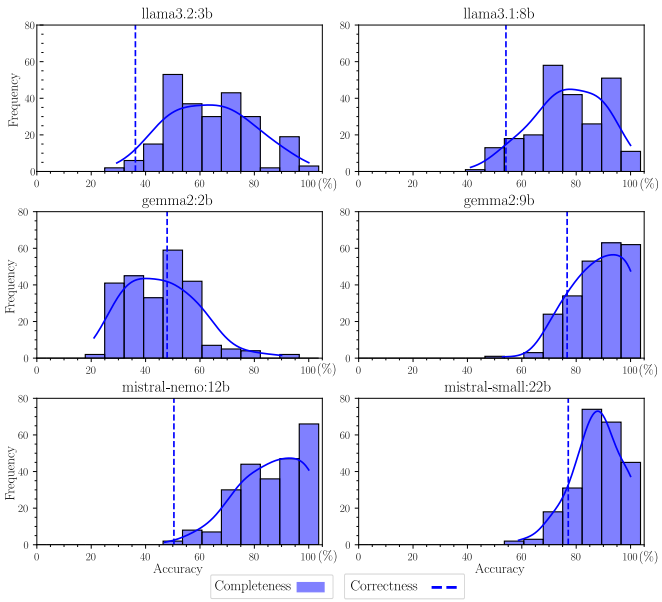

Fig. 4: Model performances on average complexity and baseline condition. Displayed metrics are correctness in % (dashed line) and completeness in % (histogram and curve). Curves are probability distributions computed through a kernel density estimation.

and that the larger the model, the better the results. However, we can notice that such a pattern occurs among model families, but not necessarily across all models, indicating that not only the model's size accounts for the performance differences. Additionally, we can note that the mistral models perform similarly to the gemma models, although they are smaller. A qualitative explanation is that the mistral models, while retrieving more concepts than the others, tended to create links between concepts which were not correlated (e.g. the robot's capability coming from the spatial property instead of solely its components). In such cases, even if the translated explanation could be considered as semantically valid, it does not respect the provided justifications.

From these results, one can observe that while model size increases, the retrieval of concepts is improved both on the number and with less spread (narrower distribution), but that not all model families report similar ranges of improvement. On the other hand, the percentage of correct explanations does not necessarily follow such a pattern, with models of different sizes exhibiting close performance.

B. Does the complexity matter? Impact of complexity level on baseline condition

As a recall, complexity is represented by the number of semantic facts in the justifications. This latter can vary given the number of axioms required to validate each atom in the rule. Consequently, one could wonder about the impact of the justifications' length on performance. Fig. 5 presents the performance of all models on the baseline condition, with each complexity level separately.

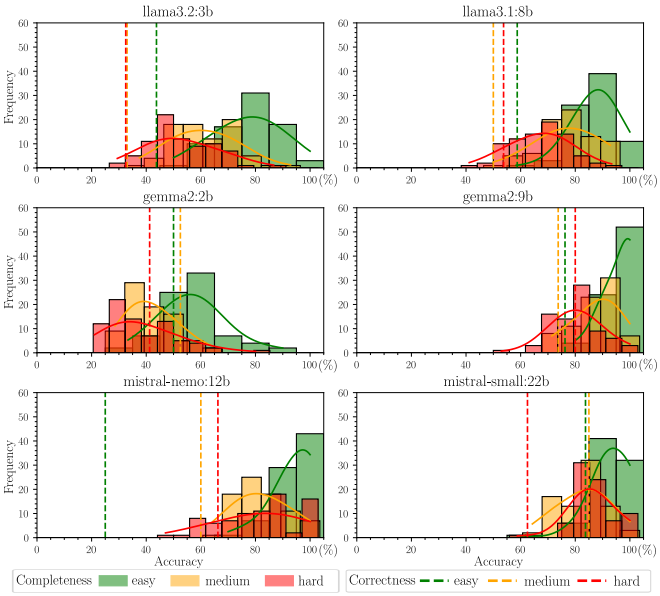

Fig. 5: Comparison between complexities (green, yellow and red) on baseline condition. Displayed metrics are correctness in % (dashed line) and completeness in % (histogram and curve). Curves are probability distributions computed through a kernel density estimation.

Completeness: The complexity level seems to have an effect on completeness performances for all models. Indeed, the harder the complexity, the more on the left the distributions are centred, indicating worse performance shown by omitting more concepts into the translated explanation with longer justification lists. This result is confirmed by the statistical tests, which show that compared to the easy complexity, the medium complexity decreases the mean completeness score by -11.9% ($p < 6.7e-14$), and the hard complexity by -18.1% ($p < 2.0e-16$). Moreover, with increased complexity the completeness distributions are not only showing lower performances, but also appear more flat, indicating more variability in the number of concepts they are able to retrieve.

Correctness: Given the correctness dashed lines on the figure, we observe that larger versions of the llama and gemma families do not seem impacted by the complexity. Indeed, the dashed lines are rather close to one another, whereas their smaller versions show a decrease in performance with hard complexity (and medium for llama3.2:3b). This result indicates that the larger versions are less sensitive to the complexity as well as performing better. However, models from the mistral family are strongly impacted by increasing complexity and not necessarily in the expected way. Indeed, the smaller version shows increasing correctness scores with increasing complexity, which is quite unexpected since one could believe that the more complex the translation, the worse performance would be exhibited. This behavior could be explained by the fact that models from this family tended to produce more compact explanations, increasing the chances that they would correlate semantic triples which were not supposed to be in the justifications.

From these results, the impact of complexity does not seem to exhibit a similar tendency across all models for the percentage of correct explanations, but does have a significant impact on the retrieval of concepts with increasing complexity leading to higher number of concepts omitted in the explanation.

C. Justification order matters? (Baseline vs Shuffle)

Since SWRL reasoners might resolve a given rule differently according to their ontology exploration method, the justifications' order could be different than the one provided in the baseline condition, which organizes each atom's justifications in a rather logical/intuitive flow. One could wonder if this factor would also influence performance, with models being sensitive to order of the justifications. A thorough understanding of the semantic facts would result in rather unchanged performance between the baseline and the shuffle conditions, since the semantic information would be the exact same. The comparison between the baseline and the shuffle is presented in Fig. 6.

Completeness: We can observe that every model is impacted by shuffling the justifications' order, especially the larger models with a larger offset between their distributions compared to their smaller versions. Indeed, the completeness distributions are more shifted to the left, while also being more flat, indicating that more concepts are missing from the explanations compared to the baseline and that the explanations show more variability. This drop in performance is significant ($p < 3.6e-10$) with a decrease of -8.9%

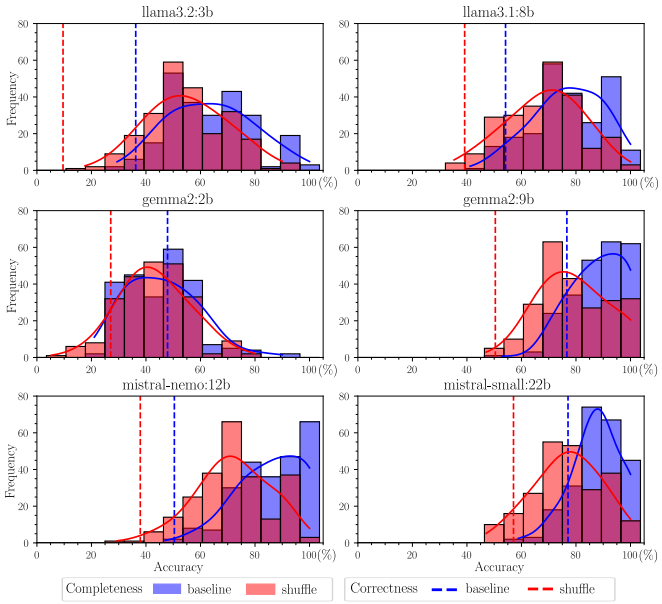

Fig. 6: Comparison between baseline (blue) and shuffle (red) conditions on average complexity. Displayed metrics are correctness in % (dashed line) and completeness in % (histogram and curve). Curves are probability distributions computed through a kernel density estimation.

on the median score, thus validating that the order of the justifications matters. Additionally, the models which appear to be the least impacted are the ones performing the worse across all models (llama3.2:3b and gemma2:2b). Thus, one could conclude that since these models only retrieved a rather small number of concepts, shuffling the order would still lead to a similarly poor performance.

Correctness: We can observe that the correctness score is also strongly impacted, with all models showing large drops in performance. Indeed, shuffling the justifications' order increases the chances that some concepts would be wrongly correlated, since this new proximity between semantic triples could be misleading, and thus lead to incorrect explanations. For instance, these contextually dependent semantic triples could lead to a property applied to an individual (e.g. the gripper's opening width), being translated as related to another individual (e.g the object's width). Compared to the baseline, the shuffle condition shows a significant decrease in performance ($p < 5.6e-6$) of -20.0%, thus validating that the order of the semantic facts leads to concepts which should not be linked together, or explanations not respecting the expected output structure.

From these results, we can conclude that the order of the justifications matters, having a significant effect on both metrics. However, we studied here the effect of random shuffling, which may not accurately reflect the output of a SWRL reasoner. Indeed, one could expect a more structured output but which would still depend over the reasoner, the rule's structure and the exploration method.

D. Adding the rule to guide the generated explanation? (Baseline vs Rule)

Since the inferences to be explained were computed by the validation of a SWRL rule, one could hypothesize that adding the rule as additional context for the model would increase performances. This hypothesis has been evaluated by comparing the baseline and the rule conditions and the results are illustrated in Fig. 7.

Completeness: The completeness distributions of all models seem to be mostly unchanged by adding the rule as additional context, indicating that the semantics in the justifications provided to the model (same for both conditions) are not significantly ($p = 0.31$) benefiting this metric's results. However, we can notice a slight drop on most of them, which could be attributed to the fact that adding this additional context tends to limit the model into incorporating concepts from the justifications, rather focusing more only on the ones provided in the rule (which doesn't contain all of the expected concepts).

Correctness: On the other hand, the correctness score is improved on almost every model (to the exception of gemma2:9b). Indeed, we can observe that the correctness dashed lines are more centred to the right, indicating better performance. This observation is supported by the statistical tests, which show a performance improvement of +10.0% ($p < 1.4e-2$) compared to the baseline. This result can be attributed to the fact that adding the rule gives the

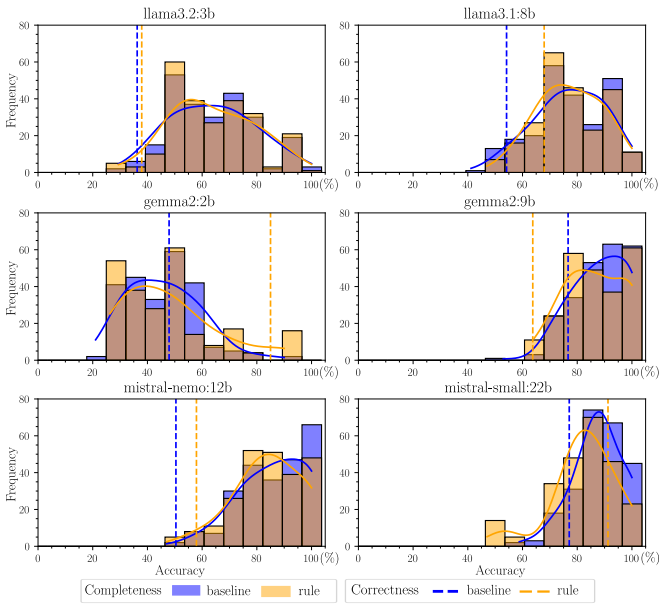

Fig. 7: Comparison between baseline (blue) and rule (yellow) conditions on average complexity. Displayed metrics are correctness in % (dashed line) and completeness in % (histogram and curve). Curves are probability distributions computed through a kernel density estimation.

model the structure of the sentence to be generated, allowing it to avoid correlating concepts which are not linked in the justifications. The most illustrative example of such a tendency is gemma:2b, which has the biggest improvement in correctness performance, with the model mostly relying on the rule and outputting a structure that matches more closely a verbalized version of the rule.

From these results, one could conclude that in order to ensure the most accurate performances, the SWRL rule should be provided in input to the LM. However, only adding the rule at the end of the prompt has been evaluated, thus, other techniques could be investigated to further improve performances.

VI. CONCLUSION

Translating inferences made in an ontology based on the justifications to natural language improves understandability for non-expert users. In this paper, we evaluated how reliable language models can be on a translation task, from ontology format to natural language sentences. The evaluation was performed on a generated dataset with action-related rule-inferred semantic triples, along with their justifications and the corresponding SWRL rules. The correctness and completeness metrics were used to assess the models' performance, respectively the number of logically correct explanations and the ratio of concepts which were incorporated into the explanation. Based on these metrics, we showed that random justification orders significantly decrease performances on both metrics, validating the hypothesis that order matters and that the model benefits from the data structure. However, we compared performances to a condition representing a

rather intuitive flow of the SWRL rule resolution, but this order is not necessarily the one showing the best performances. Thus, further investigation should be conducted in order to evaluate if a specific manner of organizing justifications would lead to better performances. We also showed that adding the used SWRL rule for an inference, as additional context, allows to significantly improves correctness, leading to more reliable results, but does not have a significant impact on completeness. Similarly to the previous argument, the manner the SWRL rule is created should follow a logical flow, since the model might understand it differently given its intuitiveness. Given the results, one should consider enhancing smaller models with techniques such as adding the rule as additional context, instead of just considering the largest models. However, while those enhanced queries would lead to more reliable explanations regarding validity, they would still remain less complete than larger ones.

Regarding the performance metrics used in this paper, it would be interesting to design a finer version of the correctness metric than just a binary metric. Indeed, while most models exhibited semantically close explanations given the justifications, such a metric does not allow to assess "how far" from a logically valid explanation the answer was. Automated annotation methods such as computing cosine similarity or maximum inner-product search would not be of help since they do not take into account the logical validity but only the semantic proximity between sentences.

This evaluation was conducted on an robotic action-oriented ontology but the results should be comparable with other ontology knowledge bases. As long as the ontology represents semantically close informations to common knowledge, we can expect a similar tendency to be shown. Since the aim of this paper was to evaluate the models in a reference case, one could now investigate multi-step reasoning and fine-tuning to assess thoroughly the performance improvement of such techniques.

TABLE I: Model performances on average complexity given the conditions. Both metrics are shown in percentages with Completeness as mean and standard deviation, and Correctness as ratio of correct explanations.

Model	**Condition**	**Correctness** (%)	**Completeness** (mean ± SD (%))
llama3.2:3b	Baseline	36.2	63.7 (±15.2)
	Shuffle	9.6	55.2 (±14.6)
	Rule	37.9	63.7 (±15.0)
llama3.1:8b	Baseline	54.2	77.2 (±12.7)
	Shuffle	39.2	68.4 (±13.8)
	Rule	67.9	77.3 (±11.7)
gemma2:2b	Baseline	47.9	45.8 (±13.1)
	Shuffle	27.1	43.2 (±12.8)
	Rule	85.0	47.9 (±17.0)
gemma2:9b	Baseline	76.7	88.1 (±9.9)
	Shuffle	50.4	78.9 (±12.6)
	Rule	63.7	86.2 (±10.6)
mistral-nemo:12b	Baseline	50.4	86.0 (±12.0)
	Shuffle	37.9	72.0 (±13.2)
	Rule	57.9	83.9 (±12.2)
mistral-small:22b	Baseline	77.1	87.5 (±8.9)
	Shuffle	57.1	77.0 (±12.1)
	Rule	91.2	82.5 (±11.5)

APPENDIX

REFERENCES

[1] J. Wang, E. Shi, H. Hu, C. Ma, Y. Liu, X. Wang, Y. Yao, X. Liu, B. Ge, and S. Zhang, "Large language models for robotics: Opportunities, challenges, and perspectives," *Journal of Automation and Intelligence*, 2024.

[2] M. Ahn, A. Brohan, N. Brown, Y. Chebotar, O. Cortes, B. David, C. Finn, C. Fu, K. Gopalakrishnan, K. Hausman *et al.*, "Do as i can, not as i say: Grounding language in robotic affordances," *arXiv preprint arXiv:2204.01691*, 2022.

[3] D. Sobrín-Hidalgo, M. A. González-Santamarta, Á. M. Guerrero-Higueras, F. J. Rodríguez-Lera, and V. Matellán-Olivera, "Explaining autonomy: Enhancing human-robot interaction through explanation generation with large language models," *arXiv preprint arXiv:2402.04206*, 2024.

[4] J. M. Beer, A. D. Fisk, and W. A. Rogers, "Toward a framework for levels of robot autonomy in human-robot interaction," *Journal of human-robot interaction*, 2014.

[5] S. Han, H. Schoelkopf, Y. Zhao, Z. Qi, M. Riddell, W. Zhou, J. Coady, D. Peng, Y. Qiao, L. Benson *et al.*, "Folio: Natural language reasoning with first-order logic," *arXiv preprint arXiv:2209.00840*, 2022.

[6] L. Tian, X. Zhou, Y.-P. Wu, W.-T. Zhou, J.-H. Zhang, and T.-S. Zhang, "Knowledge graph and knowledge reasoning: A systematic review," *Journal of Electronic Science and Technology*, 2022.

[7] B. Jiang, Y. Xie, Z. Hao, X. Wang, T. Mallick, W. J. Su, C. J. Taylor, and D. Roth, "A peek into token bias: Large language models are not yet genuine reasoners," *arXiv preprint arXiv:2406.11050*, 2024.

[8] I. Mirzadeh, K. Alizadeh, H. Shahrokhi, O. Tuzel, S. Bengio, and M. Farajtabar, "Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models," *arXiv preprint arXiv:2410.05229*, 2024.

[9] R. Schaeffer, B. Miranda, and S. Koyejo, "Are emergent abilities of large language models a mirage?" *Advances in Neural Information Processing Systems*, 2023.

[10] X. Chen, R. A. Chi, X. Wang, and D. Zhou, "Premise order matters in reasoning with large language models," *arXiv preprint arXiv:2402.08939*, 2024.

[11] W. Sun, C. Zhang, X. Zhang, X. Yu, Z. Huang, P. Chen, H. Xu, S. He, J. Zhao, and K. Liu, "Beyond instruction following: Evaluating inferential rule following of large language models," *arXiv preprint arXiv:2407.08440*, 2024.

[12] N. E. Fuchs, K. Kaljurand, and T. Kuhn, "Attempto controlled english for knowledge representation," *Reasoning Web: International Summer School*, 2008.

[13] A. Cregan, R. Schwitter, T. Meyer *et al.*, "Sydney owl syntax-towards a controlled natural language syntax for owl 1.1." in *OWLED*, 2007.

[14] K. Kaljurand and N. E. Fuchs, "Verbalizing owl in attempto controlled english," 2007.

[15] R. Power, S. Williams, and A. Third, "SWAT Ontology Verbaliser," Tech. Rep., 2017.

[16] I. Androutsopoulos, G. Lampouras, and D. Galanis, "Generating natural language descriptions from owl ontologies: the naturalowl system," *Journal of Artificial Intelligence Research*, 2013.

[17] V. Ellampallil Venugopal and P. S. Kumar, "Verbalizing but not just verbatim translations of ontology axioms," in *Artificial Intelligence and Machine Learning: Benelux Conference on Artificial Intelligence*. Springer, 2022.

[18] X. Hao, L. Cui, C. Tao, K. Roberts, and M. Amith, "Analyzing llama 3-based approach for axiom translation from ontologies," in *CEUR Workshop Proceedings*. CEUR-WS, 2024.

[19] A. Zaitoun, T. Sagi, and M. Peleg, "Generating ontology-learning training-data through verbalization," in *Proceedings of the AAAI Symposium Series*, 2024.

[20] T. Nguyen, R. Power, P. Piwek, and S. Williams, "Measuring the understandability of deduction rules for owl," 2012.

[21] M. R. Schiller, F. Schiller, and B. Glimm, "Testing the adequacy of automated explanations of el subsumptions." *Description Logics*, 2017.