

# VERUS-LM: a Versatile Framework for Combining LLMs with Symbolic Reasoning

Benjamin Callewaert

Dept. of Computer Science  
De Nayer Campus, KU Leuven, Belgium  
Leuven.AI – KU Leuven Institute for AI

Flanders Make – DTAI-FET

{benjamin.callewaert, s.vandeveldel}@kuleuven.be

Simon Vandeveldel

Joost Vennekens

Vrije Universiteit Brussel, Brussels, Belgium  
joost.vennekens@vub.be

A recent approach to neurosymbolic reasoning is to explicitly combine the strengths of large language models (LLMs) and symbolic solvers to tackle complex reasoning tasks. However, current approaches face significant limitations, including poor generalizability due to task-specific prompts, inefficiencies caused by the lack of separation between knowledge and queries, and restricted inferential capabilities. These shortcomings hinder their scalability and applicability across diverse domains. In this paper, we introduce VERUS-LM, a novel framework designed to address these challenges. VERUS-LM employs a generic prompting mechanism, clearly separates domain knowledge from queries, and supports a wide range of different logical reasoning tasks. This framework enhances adaptability, reduces computational cost, and allows for richer forms of reasoning, such as optimization and constraint satisfaction. We show that our approach succeeds in diverse reasoning on a novel dataset, markedly outperforming LLMs. Additionally, our system achieves competitive results on common reasoning benchmarks when compared to similar state-of-the-art approaches, and significantly surpasses them on the difficult AR-LSAT dataset. By pushing the boundaries of hybrid reasoning, VERUS-LM represents a significant step towards more versatile neurosymbolic AI systems. All code and datasets required to reproduce the results in this text are available online: <https://gitlab.com/EAVISE/bca/verus-lm>

## 1 Introduction

Logical reasoning is an essential aspect of problem-solving, decision-making, and critical thinking, making it an important goal in Artificial Intelligence. While recent large language models (LLMs) such as GPT-4 [2] and Gemini [25] have shown an impressive leap in reasoning-like capabilities through techniques like chain-of-thought prompting, they often lack the ability to guarantee faithful and transparent reasoning. Indeed, at a fundamental level, LLMs operate as black-box probabilistic models, which makes it difficult to ensure accurate and coherent reasoning steps. As a consequence, whether they are capable of “true reasoning” is still subject of much debate [4, 30, 22].

Symbolic inference engines, on the other hand, can derive conclusions that are provably correct and typically also explainable. However, these systems struggle to interpret ambiguous natural language input. This creates an opportunity for neurosymbolic methods to combine the strengths of LLMs with symbolic reasoning systems. Interestingly, this method loosely parallels Kahneman’s theory on human cognition, in which reasoning is divided into fast and unconscious thinking (System 1) and slow and methodical thinking (System 2) [15]. Such approaches have already demonstrated impressive improvements on common reasoning datasets [18, 14, 17].

Despite these promising results, current state-of-the-art approaches have several limitations that hinder their applicability across diverse domains. One significant drawback is their heavy reliance on

task-specific prompts to generate symbolic representations, via in-context learning. This makes these approaches less general and harder to adapt to new domains, since this may require manual prompt engineering. In real-world scenarios, where problems are highly variable, task-specific prompts can cause a severe bottleneck. Furthermore, they typically try to reduce every user query to the same logical reasoning task (or to one of only a very limited set of reasoning tasks). In many domains, users have widely different kinds of queries, which correspond naturally to equally wide range of logical reasoning tasks (e.g., optimization, generating solutions, deriving consequences, and generating explanations). Therefore, there is a lack of *versatility* in the types of queries they are able to answer, limiting their effectiveness in complex domains where multiple forms of reasoning are required.

In this paper, we propose an enhanced neurosymbolic framework that addresses these limitations titles VERUS-LM. It features a generic prompting pipeline, a semantic refinement step, a clear separation between declarative knowledge from questions, and support for a wide range of logical reasoning tasks. In this way, it is able to handle complex and dynamic environments in a robust and effective manner.

## 2 Related Work

Improving the reasoning capabilities of LLMs has recently become a topic of great interest. For the purpose of this paper, however, we will not discuss works on improving LLM reasoning with in-context prompting or by fine-tuning, but refer to [13, 19] for more information on those techniques. Instead, we will focus on the neurosymbolic approach in which an LLM is coupled with an explicit symbolic reasoning engine.

Current state-of-the-art systems all largely work in the same way. Given a reasoning problem, which typically consists of a description of some domain and a query that needs to be answered, they first let an LLM generate a formal representation of the problem and then pass this on to a reasoning engine to solve. In this way, they combine the benefits of both kinds of AI systems: the LLM offers a user-friendly natural language interface, while the reasoning engine ensures correct reasoning. This offers a promising path towards improving the reasoning skills of LLM-based systems. We will now briefly go over some of these systems and touch on how they differ.

In [14], the authors present a pipeline to automatically generate Answer Set Programming (ASP) representations [5]. Their method uses the LLM three times: first to extract relevant constants, next to generate predicates, and finally to generate the ASP rules. Using a logic puzzle dataset, they demonstrate a 71% increase in accuracy compared to a baseline GPT-4. The same authors also rely on ASP in the [LLM]+ASP framework [32], where given a problem description, the LLM extracts a set of atomic facts. These facts, coupled with pre-defined, hand-crafted “knowledge modules” containing domain-specific ASP rules, are fed to an ASP solver to find a solution. They validate their approach on four “story-based” datasets, on which it outperforms the state-of-the-art.

In the LINC system [17], an LLM is used to generate statements in First Order Logic (FOL), which are given to a FOL reasoner. To mitigate formalization errors, they generate  $K$  formalizations and make use of  $K$ -way majority voting to decide on the correct response. They demonstrate that their approach is competitive on the FOLIO dataset [12], and achieves remarkable improvements on the ProofWriter dataset [24].

Finally, though similar to the previous approaches, Logic-LM [18] has two distinctive features. First, instead of relying on a single formalization language, their system supports logic programs, FOL, constraint satisfaction problems and SAT encodings. In this way, the system is able to handle a broader range of problem types. Second, they introduce a *self-refinement stage*, in which the LLM iteratively attempts

to fix any syntax errors based on feedback from the solver. They validate their approach on five datasets, on which they generally outperform GPT-4 using chain-of-thought as baseline.

It is worth noting that most of the aforementioned methods rely heavily on dataset-specific examples, utilized for in-context learning, to guide the LLMs. Furthermore, they only support a limited number of logical reasoning tasks, which limits their generalization capabilities to new datasets. This is especially apparent on the AR-LSAT dataset [33], featuring logical reasoning questions from the Law School Admission Test, which is high in diversity and thus requires broad reasoning capabilities. For instance, given knowledge on the types of CDs sold by a store, the questions can range from “Which CDs are on sale?” to “Given  $\phi$ , which statements are true?” and “What are the minimum CDs required to be on sale if  $\phi$ ?”, for some additional statement  $\phi$ . Though it still improves on the baseline GPT-4, Logic-LM only achieves a 43.04% accuracy on this dataset due to its diversity.

Though different from the others, we also briefly discuss the SymbCoT framework [31] for the sake of comparison. Like the aforementioned works, SymbCoT internally translates a problem into FOL using in-context learning. However, instead of using a separate reasoning engine, the LLM itself then performs logical reasoning on the generated statements in a step-by-step way similar to Chain-of-Thought, instructed by specific prompts that try to mimic the kinds of logical reasoning found in logical solvers. By not using a separate reasoning engine, the approach becomes more robust to syntax errors. Interestingly, this approach consistently outperforms standard Chain-of-Thought and Logic-LM on the PrOntoQA [21], FOLIO and ProofWriter datasets.

### 3 System Design

#### 3.1 Requirements

As discussed in the previous section, current state-of-the-art approaches typically focus on a single reasoning tasks. However, more realistic use cases will likely require a more versatile approach. Imagine, for instance, a chat bot acting as a digital assistant in the field of car insurance. A person interacting with such an AI system will not only ask questions like “Am I eligible for an insurance policy?”, but will also ask other things such as “Why am I (in)eligible?”, “Depending on my car type, what would my insurance premium?” and “What elements can I change to minimize my premium?”.

For a system to be capable of handling such use cases, there are two main requirements. First, the system must support multiple modes of reasoning, which means that reasoning-specific LLM prompts will not work well (as evidenced by the results of state-of-the-art systems on the AR-LSAT dataset). Second, the reasoning component of such a system must be capable of performing different kinds of reasoning, either using a single versatile engine or, similar to Logic-LM, using multiple ones. However, this latter approach has the downside that the domain knowledge needs to be formalized anew for each type of reasoning, increasing the computational cost and risk of errors. We therefore prefer the former approach of using a single reasoning engine that supports different forms of reasoning.

Given this, we have the following design requirements for VERUS-LM:

1. *Versatile*: it should support multiple *useful* forms of reasoning (e.g., verification, explanation, optimisation, etc.)
2. *Knowledge reuse*: it should distinguish between “domain knowledge” and “task to be performed”, allowing the domain knowledge to be reused for different tasks.
3. *Generic*: all aspects of the tool, including the prompting method, should be independent of the type of reasoning tasks or the kind of domain that is being considered.

### 3.2 Reasoning Engine

While the aforementioned requirements pertain to the general design of our system, they also constrain our selection of reasoning engine. To serve as the main reasoning core of VERUS-LM, we have selected the IDP-Z3 system [6]. IDP-Z3 is a reasoning engine for  $\text{FO}(\cdot)$ , a rich extension of classical First-Order Logic with useful features such as types, aggregates, (inductive) definitions, and more. It explicitly supports the Knowledge Base Paradigm [8], in which the same domain knowledge can be used for many different reasoning tasks.

We will briefly illustrate  $\text{FO}(\cdot)$  and the forms of reasoning supported by IDP-Z3 through a small example, referring to [6] for a more extensive overview.

As in normal FOL, a *vocabulary* in  $\text{FO}(\cdot)$  consists of a set of symbols. Because  $\text{FO}(\cdot)$  is a typed logic, this includes types, in addition to predicate and function symbols. For instance, in the below example about car insurance, we use types `Customer` and `Car`. The function `risk_factor` maps each `Car` to its associated risk factor ( $\in \mathbb{R}$ ), while the function `age` maps each `Applicant` to their age ( $\in \mathbb{Z}$ ). A number of constants (= 0-ary functions) represent the car's value, its type, and the insurance premium. Finally, a unary predicate `applicant` represents which of the customers is requesting the insurance. Again following FOL, a *structure* for a vocabulary provides an interpretation for some (not necessarily all) of the symbols in this vocabulary. Finally, a *theory* consists of a set of logical sentences. In this simple example, we have two sentences: one states that every applicant must be an adult (L18), and one defines the calculation of the insurance premium (L19).

```

1 vocabulary V {
2   type Customer := {Ann, Brit}
3   type Car := {Sedan, SUV, ...}
4   risk_factor: Car → Real
5   age: Applicant → Int
6   car_value: → Real
7   premium: → Real
8   car_type: → Car
9   premium: → Real
10  applicant: Customer → Bool
11 }
12
13 structure S:V {
14   age := {Ann → 16, Brit → 32}.
15   risk_factor := {Sedan→1.03, Truck→1.15, ...}.
16 }
17
18 theory T:V {
19    $\forall p \text{ in Customer: } \text{applicant}(p) \Rightarrow \text{age}(p) \geq 18.$ 
20    $\text{premium}() = (\text{car\_value}()/100) \times \text{risk\_factor}(\text{car\_type}()).$ 
21 }
```

As usual, we denote the value of a symbol  $\sigma$  in a structure  $S$  by  $\sigma^S$  and extend this notation to terms and formulas (e.g.,  $\text{age}(\text{Ann})^S = 16$ ). Also as usual, a structure  $S$  such that  $\phi^S = \text{true}$  for all  $\phi \in T$  is called a (*logical*) *model* of  $T$  and this is written as  $S \models \phi$ .

Given such a KB, IDP-Z3 can perform different forms of reasoning to answers different questions.

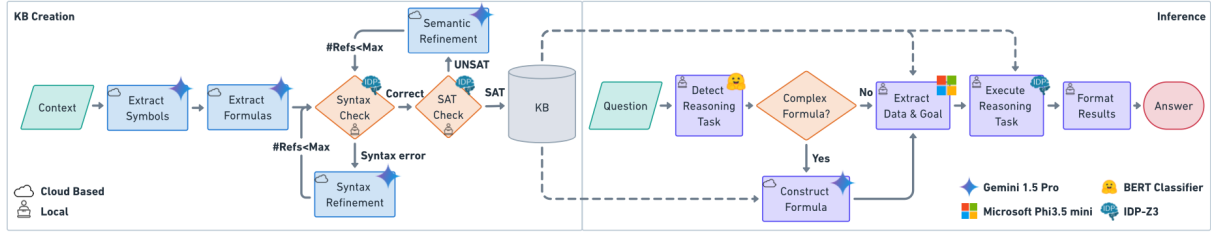


Figure 1: Flowchart depicting steps of VERUS-LM framework and the used system in each step

For instance, IDP-Z3 can determine which customers are eligible for an insurance policy, and explain why. Given the car type and car value, it can also calculate the insurance premium. Or, it can “optimize” the premium by looking for the car type with the lowest risk factor. Using these different reasoning tasks does not require any modifications to the KB itself, allowing for straightforward knowledge reuse.

These properties of IDP-Z3 make it a natural choice for our reasoning system. Additionally,  $\text{FO}(\cdot)$  is expressive enough to model many problem domains, as supported by real-life use cases [3, 9, 26, 27]. Furthermore, as First-Order Logic is the subject of many articles and textbooks, the training corpora of LLMs will also contain enough information to allow them to be fluent in this formalism.

### 3.3 VERUS-LM Architecture

The VERUS-LM framework integrates natural language processing and formal reasoning through a structured, two-phase process, as depicted in Figure 1. The first phase, Knowledge Base Creation, uses an LLM to translate domain knowledge into a symbolic  $\text{FO}(\cdot)$  specification. Note that, in this phase, we do not yet consider specific questions and instead create a reusable KB. Questions are only considered in the second phase, the Inference phase, where they are interpreted and then answered by means of an appropriate call to one of IDP-Z3’s reasoning tasks. In other words, when given multiple questions about the same domain it is only formalized once, greatly reducing the computational cost.

## 4 Knowledge Base Creation

The first phase of the VERUS-LM framework consists of transforming domain knowledge into a formal KB. This translation is performed in three steps, which we will discuss in the following sections.

### 4.1 Symbol Extraction

The LLM first identifies the relevant types, functions, and predicate symbols for the problem domain. We use a prompt that describes these three different kinds of symbols and how they are typically used. Important here is that we do not include examples from a specific dataset. Instead, our prompt just provides general instructions on how to identify and categorize the different concepts, and provides three general examples that do not appear in any of the datasets as an illustration. The prompt also contains the instruction to annotate each symbol with its informal language meaning, so that this information is present in the next phases of the pipeline.

One limitation of IDP-Z3 is that it does not support reasoning under the Open World Assumption, which is necessary for the FOLIO and Proofwriter datasets. Instead, IDP-Z3 always makes the Closed

Word Assumption (CWA). In particular, it requires that the extension of all types (apart from built-in types  $\mathbb{R}$  and  $\mathbb{Z}$ ) is always fully enumerated. As a work-around, VERUS-LM simulates OWA reasoning by introducing an additional “unknown” domain element into each type. Even though this is not theoretically correct, it persistently yields the correct results in practice, as shown in the experiments described further on.

## 4.2 Formula Extraction

During the formula extraction step, an LLM generates an  $\text{FO}(\cdot)$  theory that corresponds to the description of the problem domain. To guide this generation process, a brief overview of the  $\text{FO}(\cdot)$  grammar and clear instructions are presented, supported by two illustrative examples. These examples adopt a step-by-step approach, pairing each logical sentence with its corresponding natural language meaning for clarity. Additionally, the LLM is instructed to also consider implicit and commonsense knowledge, demonstrated in the examples. At the end of the prompt we add the previously generated vocabulary, along with the natural language description of the problem to formalize.

## 4.3 Self-Refinement

Although in our experience the LLM often produces correct logical sentences, errors still occur. To correct these, we introduce a self-refinement step that uses feedback from the reasoning engine to correct erroneous logical statements. This approach is inspired by recent work on imperative programming [7, 16] and the Logic-LM framework [18]. However, whereas these approaches only consider syntactic correctness, we introduce a novel second step to take semantic correctness into account as well, by means of a satisfiability check. These refinements are part of an iterative process, where VERUS-LM will repeat them until no errors are found or a maximum of attempts is reached.

**Syntax Refinement** If the reasoning engine finds a syntax error, the LLM is instructed to correct its output, i.e., either the vocabulary and/or the theory it produces. To this end, the LLM is prompted with its previous erroneous output, the solver’s error message, and some examples of common syntax errors and their remedies. Detailed error descriptions are provided by FOLINT [28], IDP-Z3’s code analysis tool.

**Semantic Refinement** Once the KB is syntactically correct, we use a satisfiability check to find additional semantic errors. It is reasonable to assume that the KB should be satisfiable, because, at this point, it contains only a formalization of the provided domain knowledge. For instance, if we were solving a planning problem, the KB would contain a description of the actions that can be executed, but we would not yet include a specific goal that the plan should achieve. If this KB is already inconsistent, then we take this as a sign that the LLM made a mistake in translating the domain knowledge and we instruct it to refine its previous output, based on an explanation of why this was unsatisfiable. IDP-Z3 can give such an explanation itself, by generating a “minimal unsatisfiable subset” of conflicting assignments and constraints [6].

## 5 Inference

Once the KB is constructed, the inference phase attempts to correctly answer the question. To this end, VERUS-LM follows a series of steps, as illustrated on right side of Figure 1. Given a question,

the framework starts by identifying the intended reasoning task and extracting the relevant information, which are passed on to the reasoning engine. Once the latter has finished, its output is formatted back to natural language.

## 5.1 Detect Reasoning Task

First, VERUS-LM classifies the user’s question into one of the following eight forms of reasoning.

1. **Model Generation/Expansion:** generate  $n$  logical models of the theory  $T$  (i.e., structures  $S$  such that  $S \models T$ ). Possibly, the interpretation  $\sigma^S$  that some of the symbols  $\sigma$  should have in this model  $S$  is already given (e.g., we might provide the value for constants `car_value` and `car_type` and ask the reasoner to complete the interpretation for the remaining symbols).
2. **Satisfiability:** verify if at least one model  $S$  exists for a given theory  $T$ .
3. **Optimization:** find the model  $S$  of a given  $T$  in which a given term  $t$  reaches its minimal/maximal value  $t^S$ .
4. **Propagation:** determine which atomic formulas are true / false in *all* models  $S$  of the given theory  $T$ .
5. **Explain:** explain why a given atomic formula is true / false in all models of  $T$ , or, in  $T$  has no models, then explain the inconsistency.
6. **Determine Range:** determine the range of possible values for a given function term  $f(\vec{t})$  given a theory  $T$ , i.e., the set of all values  $v$  such that there exists at least one model  $S$  of  $T$  in which  $f^S(\vec{t}^S) = v$ .
7. **Relevance:** determine which symbols  $\sigma$  are relevant, in the sense that there exists a model  $S$  of  $T$  such that  $S$  would no longer be a model if the value of  $\sigma$  in  $S$  were different.
8. **Logical entailment:** verify whether a given statement  $\phi$  is logically entailed by theory  $T$ .

All of these were either already directly offered by IDP-Z3 (1-7) or trivial to implement (8). This demonstrates that IDP-Z3 indeed provides the versatility that is required by the VERUS-LM framework.

Detecting the correct inference to be executed is crucial to ensure that questions are correctly answered. To detect the intended inference from the a question, we fine-tuned a BERT sentence classifier [20] on a custom dataset, which contains over 1000 natural language questions and their corresponding logical reasoning task.

To address more complex nested questions that demand a sequence of logical reasoning steps, we developed a multi-step reasoning mode. Similar to Chain-of-Thought prompting, this mode first breaks the complex question into a series of steps. Each step is then treated as an independent reasoning task, after which the results are combined.

## 5.2 Information Extraction

Several of the forms of reasoning listed above require some additional information, in addition to a theory  $T$ . We distinguish two cases.

**Extracting data and goal detection** Typically, in addition to a theory  $T$ , an interpretation is given for at least some of the symbols of  $T$ . For instance, the problem statement may be: “Given that Ann is 16 and Brit is 32, who is eligible for insurance?” Such information should be extracted and included as part of a structure on which the reasoning is performed (in this case,  $age^S = \{\text{Ann} \mapsto 16, \text{Brit} \mapsto 32\}$ ). In addition, some other reasoning tasks also require additional information of similar complexity, e.g., optimization requires that we know which term  $t$  should be minimized / maximized.

This information extraction task is made significantly easier by the fact that, at this point, we already know the vocabulary. We can therefore use language model grammars, such as llama.cpp GBNF grammar [10], to force the output to adhere to an automatically generated KB-specific grammar, ensuring that only correctly typed interpretations are generated. As we will show, this pruning of the output space allows us to achieve the same performance for this task with a small language model (SLM) as with a larger LLM. Since such a SLM can run on a standard laptop, this may significantly reduce computational costs, as well as enhance data security, by avoiding the transfer of potentially sensitive information to external servers.

**Construct complex formulas** In some case, such as for the logical entailment reasoning task, more complex additional information is required at inference time (e.g., the formula  $\phi$  for which it should be checked whether  $T \models \phi$ ). Constructing a complex formula  $\phi$  from the natural language question is of similar complexity as constructing the knowledge base in the first phase. Therefore, SLMs are ill-suited for this task, and a larger LLM should be used instead. Similarly to the KB creation step, this LLM not only generates formulas but may also add new symbols to the vocabulary.

We currently do not attempt to automatically detect whether a specific question contains complex formulas or not, and thus requires an LLM, because we may wish to make this decision depending on the application. If we do not care about computational cost, we can just always invoke the LLM for optimal accuracy. Alternatively, we could base the decision whether to use an SLM or LLM purely on the detected reasoning task (e.g., Entailment is typically more likely to need an LLM), or we could even train a specific classifier to make the decision.

## 6 Experiments

### 6.1 Validation

We first conduct a basic validation of VERUS-LM by checking (1) whether our use of a reasoning engine indeed leads to better performance on logical reasoning tasks than a stand-alone LLMs and (2) that the VERUS-LM pipeline is indeed able to correctly identify our eight different reasoning tasks from natural language text. Since existing datasets typically cover only one or at most a few reasoning tasks, we constructed our own dataset called *DivLR* in which all eight reasoning tasks are present. It consists of 115 questions covering six domains: investment Strategies, COVID restrictions, water Irrigation needs, a Handyman and two variants in the Body Mass Index (BMI) domain. The reason for having two BMI variants is that information about BMI (i.e., how to calculate it, the different risk levels associated to BMI values, etc.) probably occurs in LLM training data. In addition to our BMI domain which contains the well-known definition, we also defined a **B\*** domain in which novel formulas and ranges for BMI were made up, similar to the experiment in [11].

We compare VERUS-LM to baseline language models on this dataset. We compare with both an SLM and an LLM, and consider two versions of our pipeline: one in which our information extraction



Table 1: Results of VERUS-LM with SLM and LLM for information extraction and SLM and LLM separately as baseline. SLM = Phi 3.5 mini instruct, LLM = Gemini 1.5 Pro

System	C	S	H	I	B	B*	Avg
V-SLM	<b>94.7</b>	86.4	86.7	86.4	61.1	66.7	80.3
V-LLM	89.5	<b>100</b>	<b>93.3</b>	<b>95.5</b>	<b>88.9</b>	<b>83.3</b>	<b>91.8</b>
SLM	47.4	50	53.3	36.4	38.9	11.1	39.5
LLM	73.7	86.4	60	63.6	66.7	50	66.7

Table 2: Distribution of different reasoning tasks in custom datasets, detection accuracy of the BERT classifier and execution accuracy of VERUS-LM with an SLM (V-SLM) and LLM (V-LLM) for information extraction.

Reasoning Task	%	Detect Acc	Exec_Acc	
			V-SLM	V-LLM
Model Expansion	1.8	100	100	100
Satisfiability	9.6	81.8	100	90.9
Optimization	26.3	90.0	93.3	96.7
Propagation	28.1	76.5	68.6	91.2
Explain	8.8	100	80.0	90.0
Determine Range	14.0	92.9	100	85.7
Relevance	1.8	100	100	100
Entailment	9.6	100	27.3	90.9

is done by an SLM (V-SLM) and one in which this is done by an LLM (V-LLM). the LLM is Gemini 1.5 Pro [25] and the SLM is a quantised version of Phi 3.5 [1], a 3.8-billion-parameter language model whose inference cost is about 10,000 times lower than large models like GPT-4. All experiments were performed on a Mac M2 Pro CPU with 32GB of RAM.

Table 1 shows that both V-SLM and V-LLM consistently outperform both the SLM and the LLM on their own. However, V-SLM performs significantly worse on the two BMI domains than on the other domain, and also significantly worse than V-LLM on these domains. A deeper analysis shows that the SLM typically refuses to simply extract the information that the reasoning engine needs, and instead tries to perform BMI-calculations itself. We conjecture that this is due to the BMI-related examples that are likely part of the data on which the SLM was trained.

A similar but less striking issue can be seen when comparing the results of the LLM on both the **B** and **B\*** domain: the observed drop in accuracy (66.7→50) can be explained by the fact that the LLM has memorised the correct definition of BMI from its training data, and that is unable reason with a different definition, when one is explicitly provided. This suggests that, more generally, if an LLM at some point learns out-dated facts, it is hard to correct these in the prompt, requiring retraining the network instead. We see that same issue is still somewhat observable in V-LLM (88.9→83.3), but here the reasoning engine pipeline substantially reduces the effect.

Table 2 confirms that our BERT classifier identifies the correct logical reasoning task with an average accuracy of 92.6%, performing consistently well on all tasks (with a small dip for Propagation). For the subsequent information extraction, we see that, for most of the reasoning tasks, V-LLM and S-SLM

Table 3: Overview of datasets used for comparison with state-of-the-art

Dataset	Example	For inference	# Options	Reasoning	#
PrOntoQA [21]	“Stella is not hot”	Data	2: $\top$ , $\perp$	Satisfiability	500
ProofWriter [24]	“The rabbit visits the cat”	Data	3: $\top$ , $\perp$ , ?	Propagation	600
FOLIO [12]	“Marvin is neither a human nor from Mars”	Formulas	3: $\top$ , $\perp$ , ?	Logical entailment	204
LogicalDeduction [23]	A) “Eve finished third.”, B) “Max finished third”, ...	Only Data	3-5-7	Satisfiability	300
AR-LSAT [33]	“What is the max #CDs on sale? A) two, B) six, ...”	Formulas	5	Multiple	231

Table 4: Comparison of the results of the neurosymbolic system. \*Results from original paper. <sup>†</sup>Results from [31]. The results for Logic-LM are those with self refinements included, but without the LLM fallback in case of recurring syntax error.

System	ProntoQA	Proofwriter	Folio	Logical Deduction	AR-LSAT	Average	ST Dev
VERUS-LM	95.8	93.83	78.43	88.67	<b>68.36</b>	<b>85.02</b>	<b>11.49</b>
Logic-LM*	83.20	78.80	68.55	87.63	23.40	68.32	26.09
LINC*	x	<b>98</b>	72.50	x	x	x	x
SymbCoT*	<b>99.60</b>	82.50	<b>83.33</b>	<b>93.00</b>	43.91	80.47	21.63
GPT4 <sup>†</sup>	77.40	52.67	69.11	71.33	33.33	60.77	17.86
GPT4 - CoT <sup>†</sup>	98.79	68.11	70.58	75.25	35.06	69.56	22.80

perform about equally well. The biggest difference is seen for Entailment. This is the only reasoning task in our DivLR dataset for which complex formulas need to be constructed at inference time, which V-SLM predictably struggles with (accuracy 29.6%). For the Propagation task, we also observe a difference in accuracy (-25.6%) between V-SLM and V-LLM. This is primarily due V-SLM struggling with the BMI domain, as already discussed.

Overall, we conclude that the proposed pipeline makes sense: (1) introducing a logical reasoning engine into the pipeline indeed leads to better performance than using only a language model; (2) different logical reasoning tasks can indeed be successfully identified by a simple BERT classifier; (3) all of the necessary information for the reasoning tasks can indeed be extracted with a small language model, as long as no complex formulas need to be constructed, and with the caveat that an SLM might be less robust when handling concepts that already occurred frequently in its training data.

In our following experiments, we have always used V-SLM for benchmark datasets that do not require the construction of complex formulas at inference time and V-LLM otherwise.

Table 5: The distribution of reasoning tasks detected in AR-LSAT and VERUS-LM’s accuracy

Reasoning task	Distribution (%)	Exe_Acc (%)
Optimization	6.06	71.43
Determine range	16.88	79.49
Entailment	45.89	66.04
Satisfiability	31.17	69.01

## 6.2 Comparison to state-of-the-art

Having established that the VERUS-LM pipeline makes sense, we now compare it to the state-of-the-art on a number of standard benchmarks, summarized in Table 3. PrOntoQA, ProofWriter, FOLIO, and LogicalDeduction each target a specific reasoning task, while the challenging AR-LSAT dataset spans a broader range of reasoning tasks.

We compare VERUS-LM against three LLM-based approaches and two neurosymbolic approaches for logical reasoning: (1) GPT-4 answering directly, (2) GPT-4 with Chain-of-Thought, (3) Linc [17] (only FOLIO and ProofWriter), (4) Logic-LM [18] and (5) SymbCot [31].

As shown in Table 4, all systems achieve a roughly similar performance on the first four benchmarks, with VERUS-LM scoring at most 5% worse than the best system on each, without relying on **dataset-specific prompts**. For the challenging AR-LSAT benchmark, VERUS-LM significantly outperforms all other methods, doing about **25% better** than the second best system. We believe that this is due to (1) the **generality** of our approach, which does not use task-specific examples for in-context learning; (2) **FO(·)’s expressiveness** which allows for, e.g., straightforward representations of aggregates, such as “At least three CDs are on sale”:  $\#\{c \in CD : on\_sale(c)\} \geq 3$ ; (3) our support for **different forms of reasoning**, four of which were detected by the classifier and executed, as shown in Table 5.

## 6.3 Effect of Self-Refinement

Table 6 presents the results for each dataset of: (1) VERUS-LM without performing any refinements after the initial KB creation; (2) VERUS-LM with only syntactic refinement; (3) complete VERUS-LM with also semantic refinements. In each case, we report the *execution rate*, which is the percentage of cases in which the knowledge base was syntactically correct and satisfiable. We also report the *execution accuracy*, which is the percentage of the those for which the reasoning engine returned a correct result. We also report the product of the two as “total accuracy”.

The syntactic refinement increases the execution rate by 11.2% on average. The execution accuracy stays roughly the same, showing that KBs that had to be syntactically corrected are about as likely to correctly represent the domain knowledge as KBs that were initially already syntactically correct. When we then also include our semantic refinement, there is an additional increase in execution rate of on average 10%. This concerns cases in which the original KB was unsatisfiable, e.g., because of contradictory formulas or mistakes in the typing of a function or constant. Such errors typically arise only when the natural language description is difficult to understand, which also makes them hard to rectify in a correct way, as evidenced by the drop in execution accuracy when adding the semantic refinement. However, the total accuracy still markedly increases, showing that the semantic refinement step is indeed useful.

Table 6: Execution Rate and Accuracy of VERUS-LM in different refinements scenario’s: **No** Refinements, with **Syntax** Refinements and with **Both** Syntax and Semantic Refinements

	Refs	Exe_Rate	Exe_Acc	Total_Acc
PrOntoQA	No	74	97	71.8
	Syntax	88.2 $\uparrow$ 14.2	97.3 $\uparrow$ 0.3	86.8 $\uparrow$ 15
	Both	98.2 $\uparrow$ 10	97.6 $\uparrow$ 0.3	95.8 $\uparrow$ 9
ProofWriter	No	90	95.7	86.2
	Syntax	92.3 $\uparrow$ 2.3	95.7 =	88.3 $\uparrow$ 2.1
	Both	99 $\uparrow$ 6.7	94.8 $\downarrow$ 0.9	93.8 $\uparrow$ 5.5
FOLIO	No	71.6	80.8	57.8
	Syntax	89.2 $\uparrow$ 17.6	80.6 $\downarrow$ 0.2	74 $\uparrow$ 16.2
	Both	100 $\uparrow$ 10.8	78.4 $\downarrow$ 2.2	78.4 $\uparrow$ 4.4
Logical Deduction	No	93.3	89.6	83.6
	Syntax	93.7 $\uparrow$ 0.4	89.7 $\uparrow$ 0.1	84 $\uparrow$ 0.4
	Both	99.3 $\uparrow$ 6.3	89.3 $\downarrow$ 0.4	88.7 $\uparrow$ 4.7
AR-LSAT	No	60.2	84.3	50.8
	Syntax	81.8 $\uparrow$ 21.6	78.2 $\downarrow$ 6.2	64 $\uparrow$ 13.2
	Both	98.7 $\uparrow$ 16.9	69.3 $\downarrow$ 8.9	68.4 $\uparrow$ 4.4

## 7 Limitations

Though VERUS-LM expands on the state of the art, it faces some limitations. Like the earlier papers in literature, VERUS-LM itself does not have a method for validating formula correctness. This is an inherent problem in this type of neuro-symbolic AI in general, as natural language can be ambiguous, and LLMs can always produce errors. To address this limitation, we are looking into bringing a domain expert in the loop without prior logic experience to validate the output [29].

Additionally, the expressiveness of FO( $\cdot$ ) has limitations, making it unsuitable for, e.g., representing higher-order logic or for reasoning over OWA. Though VERUS-LM’s simulated OWA performed well on the datasets discussed in this work, it remains an approximation and thus will not be effective for all situations. Similarly, our pipeline’s semantic refinement step is also not always applicable, as there could be situations (such as reasoning over fake news) where we *can* expect unsatisfiability. Due to VERUS-LM’s reliance on a logical reasoning engine, it also inherits some of its limitations, such as SAT problems in very large domains. Furthermore, though the KB-creation of VERUS-LM is based on a generic prompting pipeline, its accuracy in entirely different logical paradigms remains uncertain.

## 8 Conclusion

This paper introduced VERUS-LM, a versatile neurosymbolic framework that integrates language models with symbolic reasoning capabilities. Its two-phased approach separates Knowledge Base creation from a separate inference phase, in which multiple different reasoning tasks can be performed on the same knowledge base. This approach has the inherent advantage that multiple questions about the same domain can be answered in a computationally efficient way, by reusing the same knowledge base. Another improvement made by VERUS-LM is that we extend the state-of-the-art syntactic refinement step with a semantic refinement step, based on a satisfiability check. Our experimental analysis shows that this

indeed improves results. VERUS-LM significantly outperforms the state-of-the-art on the challenging and diverse AR-LSAT dataset, while remaining competitive on simpler benchmarks.

## References

- [1] Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl et al. (2024): *Phi-3 technical report: A highly capable language model locally on your phone*. *arXiv preprint arXiv:2404.14219*, doi:10.48550/arXiv.2404.14219.
- [2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat et al. (2023): *Gpt-4 technical report*. *arXiv preprint arXiv:2303.08774*, doi:10.48550/arXiv.2303.08774.
- [3] Bram Aerts, Marjolein Deryck & Joost Vennekens (2022): *Knowledge-based decision support for machine component design: A case study*. *Expert Systems with Applications* 187, p. 115869, doi:10.1016/j.eswa.2021.115869.
- [4] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major & Shmargaret Shmitchell (2021): *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?* In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT '21*, Association for Computing Machinery, New York, NY, USA, pp. 610–623, doi:10.1145/3442188.3445922.
- [5] Gerhard Brewka, Thomas Eiter & Mirosław Truszczyński (2011): *Answer set programming at a glance*. *Communications of the ACM* 54(12), p. 92–103, doi:10.1145/2043174.2043195.
- [6] Pierre Carbonnelle, Simon Vandevelde, Joost Vennekens & Marc Denecker (2023): *Interactive configurator with FO(.) and IDP-Z3*. *arXiv preprint arXiv:2202.00343*, doi:10.48550/arXiv.2202.00343.
- [7] Xinyun Chen, Maxwell Lin, Nathanael Schärli & Denny Zhou (2023): *Teaching large language models to self-debug*. *arXiv preprint arXiv:2304.05128*, doi:10.48550/arXiv.2304.05128.
- [8] Marc Denecker & Joost Vennekens (2008): *Building a Knowledge Base System for an Integration of Logic Programming and Classical Logic*. In Maria Garcia de la Banda & Enrico Pontelli, editors: *Logic Programming*, 5366, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 71–76, doi:10.1007/978-3-540-89982-2\_12. Available at [http://link.springer.com/10.1007/978-3-540-89982-2\\_12](http://link.springer.com/10.1007/978-3-540-89982-2_12). Series Title: Lecture Notes in Computer Science.
- [9] Marjolein Deryck, Jo Devriendt, Simon Marynissen & Joost Vennekens (2019): *Legislation in the Knowledge Base Paradigm: Interactive Decision Enactment for Registration Duties*. In: *Proceedings of the 13th IEEE Conference on Semantic Computing*, IEEE, pp. 174–177, doi:10.1109/icosc.2019.8665543.
- [10] G. Gerganov (2023): *llama.cpp: LLM inference in C/C++*. Available at <https://github.com/ggerganov/llama.cpp>. Accessed: May 30, 2023.
- [11] Alexandre Goossens, Simon Vandevelde, Jan Vanthienen & Joost Vennekens (2023): *GPT-3 for Decision Logic Modeling*. In: *Proceedings of the 17th International Rule Challenge and 7th Doctoral Consortium @ RuleML+RR 2023 Co-Located with 19th Reasoning Web Summer School (RW 2023) and 15th DecisionCAMP 2023 as Part of Declarative AI 2023*, CEUR Workshop Proceedings. Available at <https://ceur-ws.org/Vol-3485/>.

- [12] Simeng Han, Hailey Schoelkopf, Yilun Zhao, Zhenting Qi, Martin Riddell, Wenfei Zhou, James Coady, David Peng, Yujie Qiao, Luke Benson et al. (2022): *Folio: Natural language reasoning with first-order logic*. *arXiv preprint arXiv:2209.00840*, doi:10.48550/arXiv.2209.00840.
- [13] Jie Huang & Kevin Chen-Chuan Chang (2023): *Towards Reasoning in Large Language Models: A Survey*. In: *Findings of the Association for Computational Linguistics: ACL 2023*, Association for Computational Linguistics, doi:10.18653/v1/2023.findings-acl.67.
- [14] Adam Ishay, Zhun Yang & Joohyung Lee (2023): *Leveraging Large Language Models to Generate Answer Set Programs*. In: *Proceedings of the 20th International Conference on Principles of Knowledge Representation and Reasoning*, pp. 374–383, doi:10.24963/kr.2023/37.
- [15] Daniel Kahneman (2011): *Thinking, Fast and Slow*. Farrar, Straus and Giroux.
- [16] Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh & Peter Clark (2023): *Self-Refine: Iterative Refinement with Self-Feedback*. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt & S. Levine, editors: *Advances in Neural Information Processing Systems*, 36, Curran Associates, Inc., pp. 46534–46594. Available at [https://proceedings.neurips.cc/paper\\_files/paper/2023/file/91edff07232fb1b55a505a9e9f6c0ff3-Paper-Conference.pdf](https://proceedings.neurips.cc/paper_files/paper/2023/file/91edff07232fb1b55a505a9e9f6c0ff3-Paper-Conference.pdf).
- [17] Theo Olausson, Alex Gu, Ben Lipkin, Cedegao Zhang, Armando Solar-Lezama, Joshua Tenenbaum & Roger Levy (2023): *LINC: A Neurosymbolic Approach for Logical Reasoning by Combining Language Models with First-Order Logic Provers*. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, p. 5153–5176, doi:10.18653/v1/2023.emnlp-main.313.
- [18] Liangming Pan, Alon Albalak, Xinyi Wang & William Wang (2023): *Logic-LM: Empowering Large Language Models with Symbolic Solvers for Faithful Logical Reasoning*. In Houda Bouamor, Juan Pino & Kalika Bali, editors: *Findings of the Association for Computational Linguistics: EMNLP 2023*, Association for Computational Linguistics, Singapore, pp. 3806–3824, doi:10.18653/v1/2023.findings-emnlp.248. Available at <https://aclanthology.org/2023.findings-emnlp.248/>.
- [19] Aske Plaat, Annie Wong, Suzan Verberne, Joost Broekens, Niki van Stein & Thomas Back (2024): *Multi-Step Reasoning with Large Language Models, a Survey*. *arXiv preprint arXiv:2407.11511*, doi:10.48550/arXiv.2407.11511.
- [20] Nils Reimers & Iryna Gurevych (2019): *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks*. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Association for Computational Linguistics, doi:10.18653/v1/d19-1410.
- [21] Abulhair Saparov & He He (2022): *Language models are greedy reasoners: A systematic formal analysis of chain-of-thought*. *arXiv preprint arXiv:2210.01240*, doi:10.48550/arXiv.2210.01240.
- [22] Parshin Shojaei, Iman Mirzadeh, Keivan Alizadeh, Maxwell Horton, Samy Bengio & Mehrdad Farajtabar (2025): *The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity*. *arXiv preprint arXiv:2506.06941*, doi:10.48550/arXiv.2506.06941.

- [23] AaroHi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Garriga-Alonso et al. (2022): *Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models*. *arXiv preprint arXiv:2206.04615*, doi:10.48550/arXiv.2206.04615.
- [24] Oyvind Tafjord, Bhavana Dalvi & Peter Clark (2021): *ProofWriter: Generating Implications, Proofs, and Abductive Statements over Natural Language*. In Chengqing Zong, Fei Xia, Wenjie Li & Roberto Navigli, editors: *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, Association for Computational Linguistics, Online, pp. 3621–3634, doi:10.18653/v1/2021.findings-acl.317. Available at <https://aclanthology.org/2021.findings-acl.317/>.
- [25] Team Gemini, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican et al. (2023): *Gemini: a family of highly capable multimodal models*. *arXiv preprint arXiv:2312.11805*, doi:10.48550/arXiv.2312.11805.
- [26] Simon Vandevelde, Jeroen Jordens, Bart Van Doninck, Maarten Witters & Joost Vennekens (2022): *Knowledge-Based Support for Adhesive Selection*. In Georg Gottlob, Daniela Incelesan & Marco Maratea, editors: *Logic Programming and Nonmonotonic Reasoning*, Springer International Publishing, Cham, pp. 445–455, doi:10.1007/978-3-031-15707-3\_34.
- [27] Simon Vandevelde, Joost Vennekens, Jeroen Jordens, Bart Van Doninck & Maarten Witters (2024): *Knowledge-Based Support for Adhesive Selection: Will it Stick? Theory and Practice of Logic Programming* 24(3), p. 560–580, doi:10.1017/s1471068424000024.
- [28] Lars Vermeulen, Simon Vandevelde & Joost Vennekens (2022): *FOLINT: static code analysis for FO(.)*. Available at <https://gitlab.com/krr/IDP-Z3/-/tree/main/folint>.
- [29] Stijn Voet, Christian Fleiner & Simon Vandevelde (2025-04-18): *Towards Knowledge Formalization for Domain Experts using LLMs*. In: *Proceedings of HHAI 2025*.
- [30] Zhaofeng Wu, Linlu Qiu, Alexis Ross, Ekin Akyürek, Boyuan Chen, Bailin Wang, Na-joung Kim, Jacob Andreas & Yoon Kim (2024): *Reasoning or Reciting? Exploring the Capabilities and Limitations of Language Models Through Counterfactual Tasks*. *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Association for Computational Linguistics, Mexico City, Mexico, pp. 1819–1862, doi:10.18653/v1/2024.Naacl-Long.102. Available at <https://aclanthology.org/2024.naacl-long.102/>, <https://doi.org/10.18653/v1/2024.naacl-long.102>.
- [31] Jundong Xu, Hao Fei, Liangming Pan, Qian Liu, Mong-Li Lee & Wynne Hsu (2024): *Faithful Logical Reasoning via Symbolic Chain-of-Thought*. In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, p. 13326–13365, doi:10.18653/v1/2024.acl-long.720.
- [32] Zhun Yang, Adam Ishay & Joohyung Lee (2023): *Coupling Large Language Models with Logic Programming for Robust and General Reasoning from Text*. In: *Findings of the Association for Computational Linguistics: ACL 2023*, Association for Computational Linguistics, doi:10.18653/v1/2023.findings-acl.321.

- [33] Wanjun Zhong, Siyuan Wang, Duyu Tang, Zenan Xu, Daya Guo, Jiahai Wang, Jian Yin, Ming Zhou & Nan Duan (2021): *AR-LSAT: Investigating Analytical Reasoning of Text*. CoRR abs/2104.06598. arXiv:2104.06598.