A Appendix

The appendix consists of the following content:

- A.1: Dataset generation
- A.2: Usage of long axioms in the proving process
- A.3: Additional training details
- A.4: Evaluation results on datasets for $\mathcal{L}_0$, $\mathcal{L}_1$, $\mathcal{L}_2$
- A.5: Symbolic translation
- A.6: Evaluation results on fuel cells data
- A.7: Simple quality tests

A.1 Dataset Generation

We generated our synthetic dataset DELTA$_D$, using four probabilistic context-free grammars. To ensure that we will produce datasets with inference depth $\mathcal{D} > 0$, i.e., datasets resulted from some inferencing, we generated a new statement $C \sqsubseteq D$, if C either appeared in some already generated fact or the RHS of some already generated statement.

Probabilistic Context-Free Grammars Each grammar is based on some vocabulary of terms. Pool A and Pool B are defined next.

Pool A

- **Atomic Concepts:** "red", "blue", "green", "kind", "nice", "big", "cold", "young", "round", "rough", "orange", "smart", "quiet", "furry".
- **Role Names:** "likes", "loves", "eats", "chases", "admires".
- **Individual Names:** "Anne", "Bob", "Charlie", "Dave", "Erin", "Fiona", "Gary", "Harry".

Pool B

- **Atomic Concepts:** "ambitious", "confident", "creative", "determined", "enthusiastic", "innovative", "logical", "persevering".
- **Role Names:** "admires", "consults", "guides", "instructs", "leads", "mentors", "supervises", "supports".
- **Individual Names:** "Ioanna", "Dimitrios", "Eleni", "Maria", "Manolis", "Angelos", "Panos", "Anna".

To generate the KBs, we employ a random sampling technique to select a subset of individuals, roles, and atomic Concepts from the pools mentioned above. An item from each pool has the same probability of being chosen.

The PCFG for the linguistic complexity level $\mathcal{L} = 0$ is shown in Table 12, the rest can be found in the provided URL. The PCFG shown is for Pool B. The grammars for the Pool A are similar. The probabilities in the PCFGs were determined experimentally to generate appropriate KBs that would yield the desired inferences in the minimum amount of time.

KB Sizes We utilize randomized parameters to control the size of a KB, based on the target reasoning depth of the corresponding dataset. The optimal (as we have found through experimentation) predefined ranges of the subsumption axioms and facts per reasoning depth $\mathcal{D}$ are as follows:

- For $\mathcal{D} = 0$: $|sub.\ axioms| \in [3, 8]$, $|facts| \in [1, 5]$
- For $\mathcal{D} = 1$: $|sub.\ axioms| \in [3, 8]$, $|facts| \in [2, 6]$
- For $\mathcal{D} = 2$: $|sub.\ axioms| \in [3, 8]$, $|facts| \in [3, 8]$
- For $\mathcal{D} = 3$: $|sub.\ axioms| \in [4, 8]$, $|facts| \in [5, 10]$
- For $\mathcal{D} = 5$: $|sub.\ axioms| \in [6, 14]$, $|facts| \in [6, 12]$

A.2 Usage of Long Axioms in the Proving Process

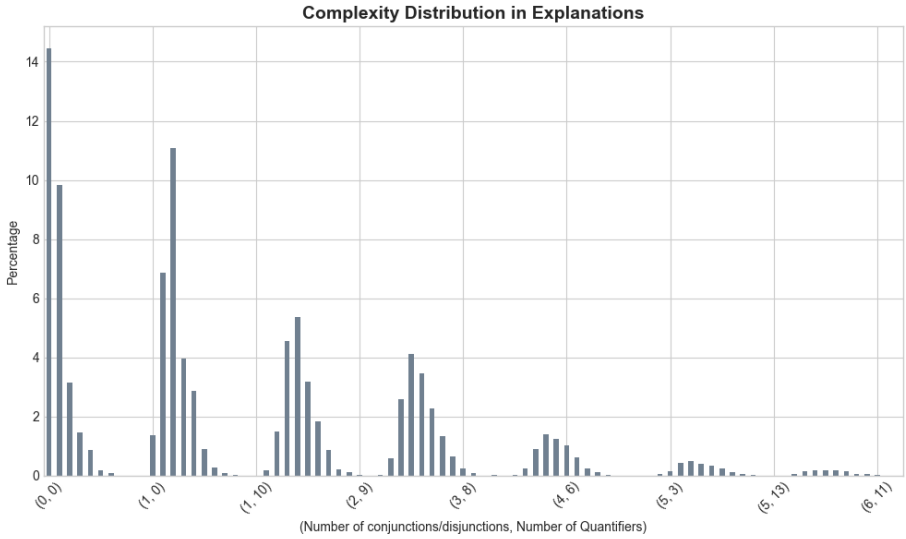

Fig. 3: Complexity distribution analysis of explanation axioms in $\mathcal{L} = 3$ KBs.

Figure 3 shows the analysis of the linguistic complexity of facts and subsumption axioms in explanations generated by $\mathcal{L} = 3$ KBs. For each axiom, we measured the number of conjunctions/disjunctions and quantifiers. The horizontal axis shows each combination of conjunctions/disjunctions and quantifiers. The vertical axis shows the frequency (as a percentage) of each combination. We observe that more than 40% of the axioms have at least two conjunctions/disjunctions *and* at least two quantifiers. Also, it is worth noting that more than 67% of the axioms have at least two conjunctions/disjunctions *or* at least two quantifiers. This indicates that long axioms are necessary, and frequently used in the reasoning process for answering the questions.

A.3 Additional Training Details

We used PyTorch 2.0 to set up our training and testing (inferencing). We use the `microsoft/deberta-v3-large` model from the transformers[11] library, along with the accelerate[12] framework.

We fine-tuned the DeBERTaV3-large model (304M parameters) using the AdamW optimizer on two A100 GPUs. We used mixed precision (FP16) for our calculations to save memory and speed up the process. The specific set of hyper-parameters used for all our models' training is given in Table 7. The model showed significant performance with this set of hyper-parameters, so there was no reason to proceed with any further hyper-parameter tuning, especially given our limited resources. The model output corresponds to the truth value 0 for `False`, 1 for `True`, and 2 for `Unknown` labels.

Table 7: Detailed specifications of the hyper-parameters used in DeBERTaV3-large training.

Hyper-parameter	**Value**
Batch size	4
Accumulation steps	2 (Effective Batch size = 8)
Learning rate	2×10^{-5}
Warm-up ratio	0.06
Epochs	4
Mixed precision	FP16
Betas	(0.9, 0.999)
Weight Decay	1×10^{-4}
Text Embedding Size	512 (dimensions)

A.4 Evaluation Results on Datasets for $\mathcal{L}_0, \mathcal{L}_1, \mathcal{L}_2$

The performance of the intermediate models DELTA$_{i,j}$, for $i \in \{0, 1, 2, 3, 5\}$, $j \in \{0, 1, 2\}$ on their corresponding datasets (of $\mathcal{D} \leq i$ and $\mathcal{L} \leq j$) are illustrated in Tables 8, 9, 10. We observe that the pattern of the models' performance across various linguistic complexity levels is similar. However, as the models progress to higher linguistic complexity levels and, hence are trained on more data, the number of times they achieve perfect accuracy is increased. The models trained on $\mathcal{D} \geq 3$ show very good generalization on unseen reasoning depths, whereas the performance on unseen reasoning depths of the models trained on $\mathcal{D} \leq 2$ fluctuates across linguistic complexity levels. This can be attributed to the complexity difference among linguistic levels, affecting models' generalization.

[11] `https://github.com/huggingface/transformers`
[12] `https://github.com/huggingface/accelerate`

Table 8: Accuracy of DELTA models on their own test sets, and the entire, and slices of $\mathcal{D} \leq 5, \mathcal{L} = 0$ dataset.

	$\text{DELTA}_{0,0}$	$\text{DELTA}_{1,0}$	$\text{DELTA}_{2,0}$	$\text{DELTA}_{3,0}$	$\text{DELTA}_{5,0}$
Test (own)	100.0	99.7	99.4	98.9	98.7
$\mathcal{D} \leq 5, \mathcal{L} = 0$	61.4	75.3	93.2	97.7	98.8
$\mathcal{D} = \text{N/A}$	98.8	97.9	94.4	95.2	98.2
$\mathcal{D} = 0$	99.9	100.0	100.0	99.5	100.0
$\mathcal{D} = 1$	48.9	99.6	100.0	99.5	100.0
$\mathcal{D} = 2$	11.9	47.1	96.3	99.0	99.0
$\mathcal{D} = 3$	34.3	49.5	75.7	99.0	99.0
$\mathcal{D} = 4$	32.1	45.0	72.0	99.0	99.0
$\mathcal{D} = 5$	29.6	42.9	67.2	97.5	99.0

Table 9: Accuracy of DELTA models on their own test sets, and the entire, and slices of $\mathcal{D} \leq 5, \mathcal{L} \leq 1$ dataset.

	$\text{DELTA}_{0,1}$	$\text{DELTA}_{1,1}$	$\text{DELTA}_{2,1}$	$\text{DELTA}_{3,1}$	$\text{DELTA}_{5,1}$
Test (own)	100.0	99.7	99.6	99.7	99.5
$\mathcal{D} \leq 5, \mathcal{L} \leq 1$	67.9	92.4	85.8	98.7	99.6
$\mathcal{D} = \text{N/A}$	99.0	98.1	99.3	98.8	99.7
$\mathcal{D} = 0$	99.9	100.0	100.0	100.0	100.0
$\mathcal{D} = 1$	52.5	99.3	99.5	100.0	100.0
$\mathcal{D} = 2$	27.4	81.7	97.5	99.0	99.5
$\mathcal{D} = 3$	47.9	79.5	61.5	99.0	99.5
$\mathcal{D} = 4$	46.9	77.2	58.0	98.0	99.0
$\mathcal{D} = 5$	39.4	66.0	53.5	96.5	99.0

Table 10: Accuracy of DELTA models on their own test sets, and the entire, and slices of $\mathcal{D} \leq 5, \mathcal{L} \leq 2$ dataset.

	$\text{DELTA}_{0,2}$	$\text{DELTA}_{1,2}$	$\text{DELTA}_{2,2}$	$\text{DELTA}_{3,2}$	$\text{DELTA}_{5,2}$
Test (own)	99.9	99.5	99.7	99.7	99.6
$\mathcal{D} \leq 5, \mathcal{L} \leq 2$	55.1	89.7	85.3	99.1	99.7
$\mathcal{D} = \text{N/A}$	99.6	95.6	98.7	99.2	99.7
$\mathcal{D} = 0$	99.9	100.0	100.0	100.0	100.0
$\mathcal{D} = 1$	37.2	99.6	100.0	100.0	99.5
$\mathcal{D} = 2$	16.1	81.7	98.5	100.0	99.0
$\mathcal{D} = 3$	16.4	62.7	59.0	98.5	99.5
$\mathcal{D} = 4$	17.6	63.1	62.0	99.5	100.0
$\mathcal{D} = 5$	10.0	63.4	51.0	98.0	98.5

A.5 Symbolic Translation

Examples of translations from natural language to both soft and hard symbolic forms are presented in Table 11. We can observe that the SoftSymbolic dataset is a good generalization test for DELTA_M as it eliminates any vocabulary influence on the model, whereas the HardSymbolic dataset offers a purely logical form of the sentences. The symbolic datasets resulted from symbolic translations of the $\mathcal{D} \leq 5, \mathcal{L} \leq 1$ test-set of DELTA_D.

Table 11: Examples of sentences' translations to soft/hard symbolic forms.

Natural language form	**Soft symbolic form**	**Hard symbolic form**
If someone mentors someone that is ambitious and that supervises less than one creative people, then they guide only people that are not persevering or that consult at most two confident people.	If someone R6 someone that is C1 and that R7 less than one C3 people, then they R3 only people that are not C8 or that R2 at most two C2 people.	exists R6 . ((+ C1) and (< 1 R7 . (+ C3))) is subsumed by only R3 . ((not C8) or (≤ 2 R2 . (+ C2)))
Maria supports less than one people that are confident or not persevering.	a4 R8 less than one people that are C2 or not C8.	(< 1 R8 . ((+ C2) or (not C8))) (a4)
If someone is not confident, then they mentor someone that is ambitious and that supervises less than one creative people.	If someone is not C2, then they R6 someone that is C1 and that R7 less than one C3 people.	not C2 is subsumed by exists R6 . ((+ C1) and (< 1 R7 . (+ C3)))

A.6 Evaluation Results on Fuel Cells Data

A sample from the fuel cells diagnostics datasets is presented in Table 13. We generated two such datasets, one involving a single sensor, and the other involving two sensors, hence making the contexts more complex. We generated these datasets using simple random sampling over predefined pools of examples so they could result in true, false, and unknown questions. We can observe that the vocabulary and the format of these data are different from the dataset DELTA$_D$ where our model was trained, although it performs particularly well (94%) zero-shot.

A.7 Simple Quality Tests

The handcrafted quality tests we created, targeting various important knowledge base equivalences, along with the predictions of DELTA$_M$ on those are presented in Table 14. Although these tests are very similar in both structure and vocabulary with the dataset that DELTA$_M$ has been trained, we can observe a blind spot of the model on cases involving numerical restrictions, tending to answer "unknown" (U) on such questions.

Table 12: Probabilistic Context-Free Grammar for $\mathcal{L} = 0$ KBs

$ABoxAssertion$	$\rightarrow ConceptAssertion \mid RoleAssertion$
$TBoxAxiom$	$\rightarrow ConceptInclusion$
$ConceptInclusion$	$\rightarrow InclusionL0$
$InclusionL0$	$\rightarrow ConceptL0 \sqsubseteq ConceptL0\,[0.6] \mid SpecialAxiom\,[0.4]$
$SpecialAxiom$	$\rightarrow {}' + ' \top \sqsubseteq \forall\, RoleName'.'('\, Concept')' \mid$ $\exists\, RoleName'.'('\, + '\, \top')' \sqsubseteq Concept$
$Concept$	$\rightarrow ConceptL0$
$ConceptL0$	$\rightarrow ConceptNameOrRestriction$
$ConceptNameOrRestriction$	$\rightarrow Polarity\, ConceptName \mid RestrictionConcept$
$RestrictionConcept$	$\rightarrow RestrictionD0$
$RestrictionD0$	$\rightarrow Restriction\, RoleName'.'('\, Polarity\, ConceptName')' \mid$ $\exists\, RoleName'.'('\, + '\, \top')' \mid \forall\, RoleName'('\, + '\, \bot')'$
$Restriction$	$\rightarrow \forall \mid \exists \mid Symbol\, Number$
$Symbol$	$\rightarrow {}' > ' \mid ' \geq ' \mid ' < ' \mid ' \leq ' \mid ' =$
$Number$	$\rightarrow '1' \mid '2' \mid '3'$
$ConceptName$	$\rightarrow 'ambitious' \mid 'confident' \mid 'creative' \mid 'determined' \mid$ $'enthusiastic' \mid 'innovative' \mid 'logical' \mid 'persevering'$
$RoleName$	$\rightarrow 'admires' \mid 'consults' \mid 'guides' \mid 'instructs' \mid$ $'leads' \mid 'mentors' \mid 'supervises' \mid 'supports'$
$IndividualName$	$\rightarrow 'Ioanna' \mid 'Dimitrios' \mid 'Eleni' \mid 'Maria' \mid$ $'Manolis' \mid 'Angelos' \mid 'Panos' \mid 'Anna'$
$RoleAssertion$	$\rightarrow RoleName'('\, IndividualName','\, IndividualName')'$
$ConceptAssertion$	$\rightarrow '('\, Concept')''('\, IndividualName')'$
$Polarity$	$\rightarrow {}' + ' \mid '\neg '$
$Connective$	$\rightarrow {}' \sqcup ' \mid '\sqcap '$

Table 13: Examples of generated <context, question, answer> triplets about fuel cells. The first example involves one sensor (s1) in its context, while the last one involves two sensors (s1 and s2).

Context	**Question**	**Answer**
s1 is a system. s1 is in a state st1. st1 is described by v1. v1 is result of an observation o1. failure mode is not a normal mode. o1 is made by vs. vs is a voltage sensor. if a system is in a state that is described by a very high voltage value that is result of an observation made by some voltage sensor that is a reliable sensor then the system is under catalyst dissolution. vs is a reliable sensor. carbon support corrosion is a failure mode. v1 is a very high voltage value.	The system is under catalyst dissolution.	True.
s1 is a system. s1 is in a state st1. st1 is described by v1. failure mode is not a normal mode. v1 is calculated by vs1 and vs2. vs1 is a hydrogen mass sensor. vs2 is a temperature sensor. if a system is in a state that is described by a very large anode humidity change that is calculated by some hydrogen relative humidity sensor that is a reliable sensor and some temperature sensor that is a reliable sensor and is described by a large cathode humidity change then the system is under membrane mechanical stress. vs1 is a reliable sensor. vs2 is a reliable sensor. catalyst dissolution is a failure mode. v1 is a very large anode humidity change that is calculated by some hydrogen relative humidity sensor that is a reliable sensor and some temperature sensor that is a reliable sensor and is described by a large cathode humidity value.	The system is under carbon support corrosion.	Unknown.

Table 14: Handcrafted quality tests for DELTA$_M$.

Context	**Question**	**Correct Ans.**	**DELTA$_M$**
Anne is red and green.	Anne is red.	T	T
	Anne is green.	T	T
Anne is red. Anne is green.	Anne is red and green.	T	**U**
If someone is blue, then they are red and green.	If someone is blue, then they are red.	T	T
	If someone is blue, then they are green.	T	T
If someone is blue, then they are red. If someone is blue, then they are green.	If someone is blue, then they are red and green.	T	T
If someone is blue, then they are red and green.	If someone is blue, then they are red or green.	T	**U**
Anne is red. Anne is green.	Anne is red or green.	T	**U**
Anne is red and green.	Anne is red or green.	T	**U**
	Anne is green or red.	T	**U**
Anne is red or green.	Anne is green or red.	T	T
If someone is blue or red then they are green.	If someone is blue, then they are green.	T	T
	If someone is red, then they are green.	T	T
If someone is blue, then they are green. If someone is red, then they are green.	If someone is blue or red, then they are green.	T	T
	If someone is blue and red, then they are green.	T	T

Continued on next page

Table 14 continued from previous page

Context	**Question**	**Correct Ans.**	**DELTA**$_M$
People that eat someone red or green, are blue.	People that eat someone red or eat someone green, are blue.	T	T
	People that eat someone red or green, they eat someone red or eat someone green.	T	T
	People that eat someone that is red or eat someone that is green they eat someone that is red or green.	T	T
Blue people eat someone red or green.	People that are blue they eat someone that is red or they eat someone that is green.	T	T
People that eat only people that are red or green are blue.	People that eat only people that are red or eat only people that are green, are blue.	T	T
People that eat something are blue. Anne eats Bob. Bob is green.	Anne is blue. Anne is green.	T U	T U
People that eat something are blue. Anne eats Bob. Bob is green. If someone is blue, then they are not green.	Anne is blue. Anne is green.	T F	T F
Someone can like only people that are nice. Bob is not nice.	Anne likes Bob.	F	**U**
Someone can like only people that are nice. Bob is nice.	Anne likes Bob.	U	U

Continued on next page

Table 14 continued from previous page

Context	**Question**	**Correct Ans.**	**DELTA$_M$**
Anne likes less than two people. Anne likes Bob. Anne likes John.	Anne likes Alice.	F	**U**
Anne likes Bob.	Anne likes none.	F	F
Anne likes Bob. Anne likes John. Anne likes Alice.	Anne likes more than two people.	T	**U**
	Anne likes more than four people.	U	U
	Anne does not like less than two people.	T	**U**
Anne likes Bob.	Anne does not like Bob.	F	F