OntoLM Tutorial | Chapter 5 of 7

ISWC BARI 2026

Complex Ontology Reasoning with LLMs

Hui Yang
Research Associate
University of Manchester

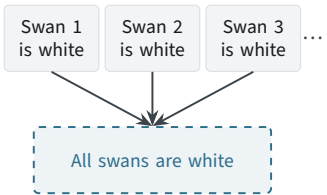
MANCHESTER 1824
The University of Manchester

University of Exeter

Three modes of reasoning

Induction

evidence ⇒ **plausible hypothesis**



probable

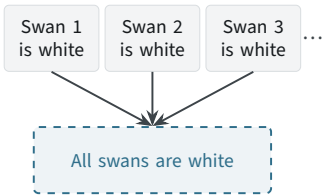
a black swan breaks it

Parts I–II: embedding methods

Three modes of reasoning

Induction

evidence ⇒ **plausible hypothesis**



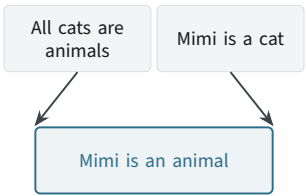
probable

a black swan breaks it

Parts I–II: embedding methods

Deduction

premises ⇒ **conclusion**



entailed

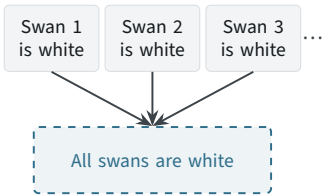
the conclusion follows if premises hold

§1: deductive benchmarks and proofs

Three modes of reasoning

Induction

evidence ⇒ **plausible hypothesis**



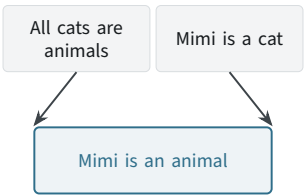
probable

a black swan breaks it

Parts I–II: embedding methods

Deduction

premises ⇒ **conclusion**



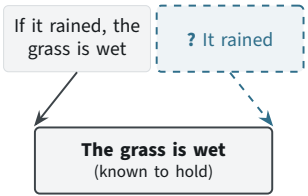
entailed

the conclusion follows if premises hold

§1: deductive benchmarks and proofs

Abduction

known fact ⇒ **missing premise**



plausible

or was it the sprinkler?

§2: NeurOWL

Deduction: entailed conclusion
Induction and abduction: plausible hypotheses

This part: what can language models do?

Deductive and abductive reasoning over ontologies

- **Deductive Reasoning** — 20 min
  Answer accuracy → proof-step grading → proof construction
- **Abductive Reasoning** — 4 min
  NeurOWL proposes a missing axiom
- **Summary** — 3 min
  Research gaps and three takeaways

1

Deductive Reasoning

Answer accuracy
Proof grading
Proof construction over ontologies

Deductive benchmarks: two distinctions

Two questions organize how we compare deductive benchmarks.

Question	One focus	Another focus
Research focus	Answer accuracy: how often is the conclusion correct?	Reasoning progress: where does the derivation succeed or fail?
What is graded?	The final answer: a label or choice.	The proof: its steps and whether they reach the goal.

Examples: OntoURL: final answer PrOntoQA: proof steps LLM4Proof: formal proof

> From answer correctness to a checkable proof

OntoURL[81]: grading the answer

OntoURL covers 15 tasks in understanding, reasoning, and learning.
Here we compare two class-membership tasks.

Toy ontology (illustrative)



U4: stated class
Which class is Nori explicitly asserted to belong to? **Penguin.**

R3: inferred class
Which further class follows from the ontology? **Bird.**

Scoring: choose a class from multiple-choice options.
Only the final answer is graded.

Qwen2.5-72B, zero-shot on OntoURL v1.1 (43 ontologies): **U4 85.6%** vs. **R3 77.2%**

R3: 8.4 points below U4 Score: final answer only

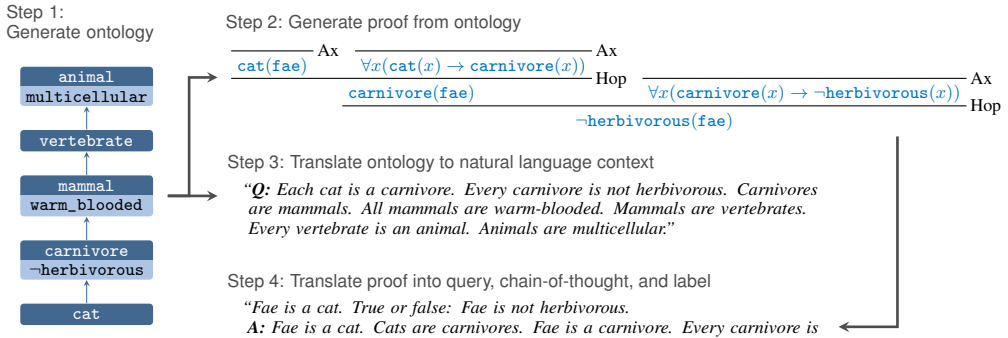
Zhang et al., J. Web Semantics 2026 [81]; OntoURL v1.1 task definitions and results

PrOntoQA[65] (1): testing LM reasoning

1. Goal: test whether LM chains of thought form valid, goal-directed proofs—beyond answer accuracy.

2. Dataset construction

Ontology → context; ontology → proof → query, CoT, label.



Saparov & He, ICLR 2023 [65], §3 and Fig. 2 (original)

PrOntoQA[65] (2): controlled factors

The dataset varies three conditions to test when the proof becomes harder.

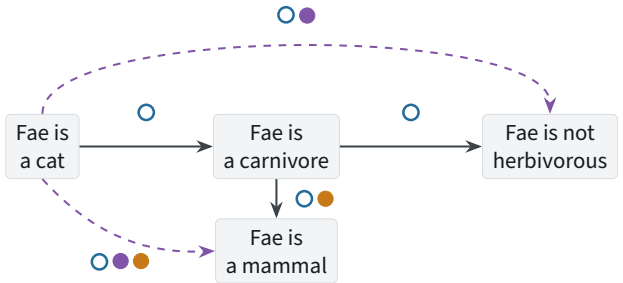
- **Proof length** 1, 3, or 5 reasoning hops
- **Context order** proof order (bottom-up) or reversed (top-down)

  Bottom-up: Cats are carnivores; carnivores are mammals; mammals are animals.

  Top-down: Mammals are animals; carnivores are mammals; cats are carnivores.
- **Prior knowledge**
  - *true*: familiar and consistent facts
  - *false*: familiar but counterfactual facts
  - *fictional*: invented names

Saparov & He, ICLR 2023 [65], §3 and §5.1; illustrative example

PrOntoQA[65] (3): grading each step of the proof



Legend

- Valid (modus ponens)
- Misleading: off path
- Skip: omitted steps
- Rule composition (broadly-valid)
- Invalid

Colors combine.

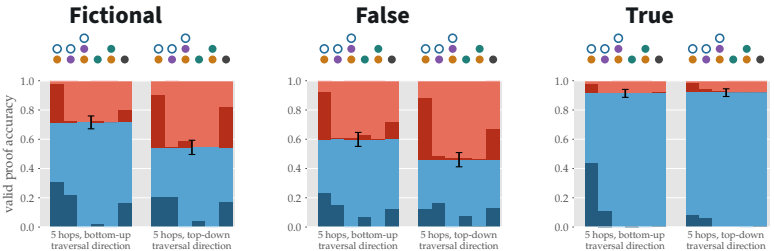
Rule composition: two rules ⇒ one new rule

Cat ⊑ Carnivore Carnivore ⊑ ¬Herbivorous
⇓ Cat ⊑ ¬Herbivorous (a rule about *all cats*)

Cases 1 2 3 4 5 6

PrOntoQA[65] (4): reading the proof analysis

InstructGPT, 8-shot CoT, 5 hops. Bar fill: blue = correct proofs; red = incorrect proofs.



Legend

- Valid (modus ponens)
- Misleading: off path
- Skip: omitted steps
- Rule composition (broadly-valid)
- Invalid

Colors combine.

Dark fill: proofs containing that step type / all examples.

Prior knowledge helps: true ontologies perform much better than false or fictional ones.
Planning remains difficult: valid local steps alone do not guarantee a successful proof.

Saparov & He, ICLR 2023 [65], Fig. 4 (original excerpts), §5.3–5.4; color markers added

Toward more complex proof

A correct answer alone does not show how the LM reached it.

What we have seen

- **Answer accuracy**: OntoURL checks the final choice
- **Short proofs**: PrOntoQA grades steps in generated linear chains

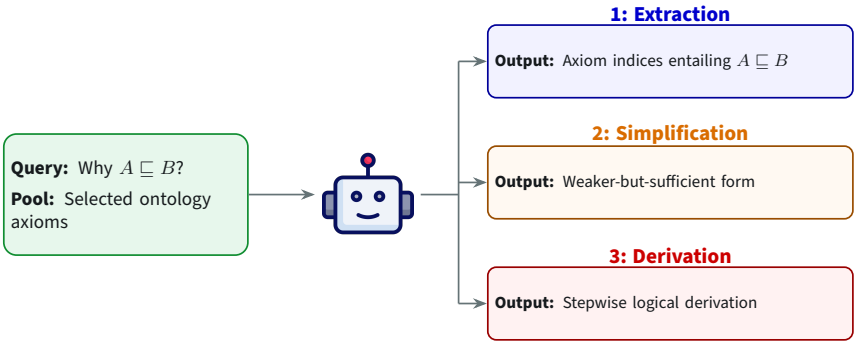
Next: LLM4Proof[60]

- Construct proofs over **real OWL ontologies**
- Study more complex, non-chain proof structure

Saparov & He, ICLR 2023 [65]; Yang et al., WWW 2026 [60]

Our focus: proof construction

Goal: Construct valid, step-by-step proofs from selected, noisy axioms in real OWL ontologies.



Yang, Chen, Sattler, WWW 2026 [60], Fig. 1

A concrete example

Premise Pool

(1) DomDog $\sqsubseteq$ Mammal
(2) Mammal $\sqsubseteq$ Animal
(3) CompAnimal $\equiv$ Animal $\sqcap$ $\exists$hasOwner.Human
(4) DomDog $\equiv$ $\exists$hasOwner.Human $\sqcap$ Dog
✗ (5) Fish $\sqsubseteq$ $\exists$livesIn.AquaticEnv
… [noise]

Prove: $\texttt{DomDog} \sqsubseteq \texttt{CompAnimal}$

Extraction

Select premises: **{1, 2, 3, 4}**

Simplification

(3′): Animal $\sqcap$ $\exists$hasOwner.Human $\sqsubseteq$ CompAnimal
(4′): DomDog $\sqsubseteq$ $\exists$hasOwner.Human

Derivation

S1: DomDog $\sqsubseteq$ Animal $\sqcap$ $\exists$hasOwner.Human
(from (1), (2), (4′))
S2: DomDog $\sqsubseteq$ CompAnimal
(from (3′), S1)

Yang, Chen, Sattler, WWW 2026 [60], Fig. 1; output format: Fig. 4

Dataset construction and evaluation

Extraction target: a justification

The minimal sufficient evidence for deriving the given conclusion.

$$J \in \mathrm{Just}(\alpha) \iff \begin{cases} J \subseteq \mathcal{O}, \quad J \models \alpha, \\ \forall J' \subsetneq J, \quad J' \not\models \alpha. \end{cases}$$

Reasoner-backed construction and evaluation

In the formal-axiom setting, we can **automatically**:

1. compute a gold justification for each selected conclusion
2. check simplified axioms and every derivation step.

- **Limited context window:** input a justification plus distractor axioms.

Yang, Chen, Sattler, WWW 2026 [60], Def. 1 and §4

Dataset construction: two difficulty measures

How hard is it to prove $\mathcal{O} \models A \sqsubseteq B$?

1. Atomic distance

$$d_{\text{sub}}(A, B) = \min\{k \mid A = A_0,\ A_k = B,\ A_0 \sqsubseteq_{\text{dir}} \cdots \sqsubseteq_{\text{dir}} A_k\}.$$

Minimum number of direct atomic subsumptions between A and B.

2. Justification length

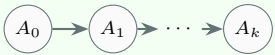
$$|J| = \#\{\text{gold axioms}\}.$$

Amount of evidence that must interact in the proof.

Difficulty: a chain versus interacting axioms

The two measures reveal different proof structures

Path-like proof



$$\mathcal{O}_1 = \{A_0 \sqsubseteq A_1, \ldots, A_{k-1} \sqsubseteq A_k\}$$

$$\mathcal{O}_1 \models A_0 \sqsubseteq A_k, \qquad d_{\mathrm{sub}}(A_0, A_k) = k.$$

$$J = \mathcal{O}_1, \qquad |J| = k.$$

Path and evidence align: close to transitive closure.

$|J| \approx d_{\mathrm{sub}} \Rightarrow$ often path-like

Non-chain proof

$$\mathcal{O}_2 = \{A \sqsubseteq \exists r.B,\ B \sqsubseteq B_1,\ \exists r.B_1 \sqsubseteq C\}$$

$$\mathcal{O}_2 \models A \sqsubseteq C, \qquad d_{\mathrm{sub}}(A, C) = 1.$$

$$J = \mathcal{O}_2, \qquad |J| = 3.$$

Concepts look close, but the proof needs extra structure.

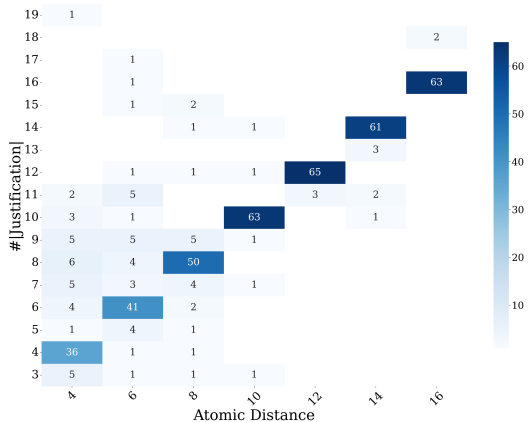
$|J| \gg d_{\mathrm{sub}} \Rightarrow$ often non-chain.

Yang, Chen, Sattler, WWW 2026 [60], §4.1 and Remark 1

Dataset construction: selected conclusions

Select conclusions across different atomic distances

- **Diagonal:** path-like proofs.
- **Upper-left:** short taxonomy path, longer proof evidence.



Selected conclusions by atomic distance and justification length.

Yang, Chen, Sattler, WWW 2026 [60], §4.3 and Fig. 2

Evaluation setup

Real ontology sources:

- Use the $\mathcal{EL}$ fragments of SNOMED CT, GO-Plus, and FoodOn.

8 LLMs Evaluated:

- **Commercial:** GPT-o4-mini.
- **32B Open:** Qwen3-32B, Qwen2.5-32B, DS-R1-Qwen-32B.
- **24B Open:** Magistral-Small-2506.
- **8B Open:** Qwen3-8B, DS-R1-Qwen-8B, DS-R1-Llama-8B.

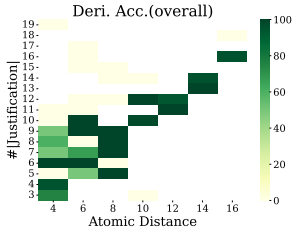
1. LLMs can prove, but struggle with complexity

GPT-o4-mini on SNOMED CT: standard vs. hard cases

Metric	Standard	Hard
Format-correct output	91.98%	51.72%
Extraction (Jaccard)	99.49%	94.39%
Simplification (whole task)	80.91%	40.14%
Valid derivation steps	92.33%	79.69%
Valid whole proofs	**80.30%**	**32.39%**

Hard: 261 further conclusions with $|J| \geq d_{\text{sub}} + 3$. Values are unweighted.

- Extraction scores the selected axioms against the gold justification.
- A whole proof succeeds only when every step is valid and the goal is reached.



Derivation accuracy by atomic distance and justification length

Yang, Chen, Sattler, WWW 2026 [60], Table 3 (App. B.1) and Fig. 7

2. Impact of Inference Rules

- **GPT-o4-mini:** Internalized logic; explicit rules can sometimes distract or hurt performance.
- **Qwen3-32B:** Benefits significantly; accuracy multiplies ($2\times$–$3\times$).

Inference Rules

```
## Inference Rules
```

1. Subsumption: If $A \sqsubseteq B$ and $B \sqsubseteq C$, then $A \sqsubseteq C$.

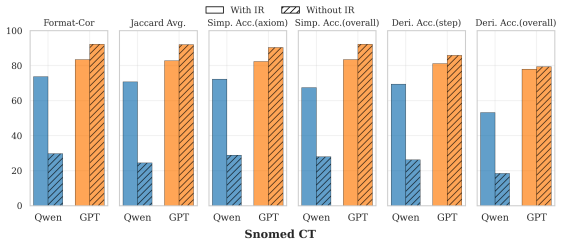
2. Conjunction: If $A \sqsubseteq B$, $A \sqsubseteq C$, then $A \sqsubseteq B \sqcap C$.

3. Existential: If $A \sqsubseteq \exists r.C$ and $C \sqsubseteq D$, then $A \sqsubseteq \exists r.D$.

4. Equivalence: If $A \equiv B$, then $A \sqsubseteq B$ and $B \sqsubseteq A$.

5. Conjunction Inference: If $A \sqsubseteq B \sqcap C$, then $A \sqsubseteq B$ and $A \sqsubseteq C$.

Here, $A, B, C, \ldots$ are concepts, r denotes relations. $\sqsubseteq$ means subclass, $\equiv$ means equivalence, $\exists r.A$ means the set of elements having a relationship r with some element in A.



Yang, Chen, Sattler, WWW 2026 [60], Fig. 5; original talk slide "Impact of Inference Rules"

3. Vulnerability to Noise

Increasing Noise Ratio (SNOMED CT):

- Evaluated ratios from 1:1 to **1:20**.

At 1:20 noise ratio (GPT-o4-mini):

- **Extraction:** drops ∼47%
- **Simplification:** drops ∼42%
- **Derivation:** drops ∼53%

Yang, Chen, Sattler, WWW 2026 [60], Table 2 and §6.2; original talk slide

4. Logical Form vs. Natural Language

🇦🇪 Does translating to English hurt?

- *Formal:* $\texttt{Teacher} \equiv \texttt{Person} \sqcap \exists \texttt{teach.Course}$
- *NL:* “Teacher is a person who teaches some courses”

Result:

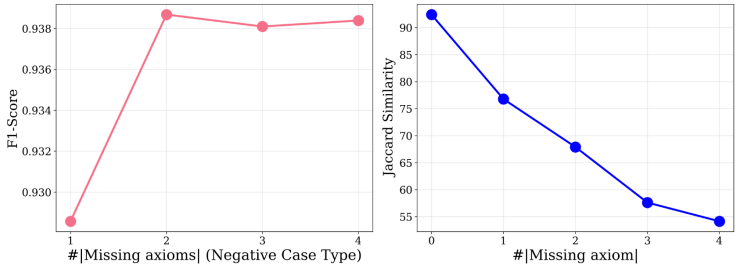
- Only $< 8\%$ performance drop.
- Logical complexity is the strict bottleneck, not format.

Yang, Chen, Sattler, WWW 2026 [60], Table 2 and §6.2; original talk slide

5. Detecting Incomplete Premises

What if some axioms are missing?

- LLMs detect the incomplete case well (F1 $\approx$ 0.93).
- Identifying the remaining useful pieces is hard.
  Jaccard similarity drops up to **38%**.



Yang, Chen, Sattler, WWW 2026 [60], Fig. 8 and §6.3; original talk slide

2

Abductive Reasoning

The missing-axiom problem
NeurOWL: proposing a plausible bridge

Abduction over ontologies: the problem

Problem setting

Input
An incomplete ontology $\mathcal{O}'$ and a query α, with $\mathcal{O}' \not\models \alpha$.

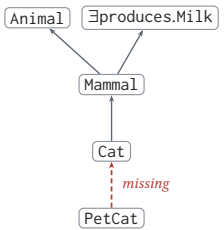
1. Judge the query
Should α hold? Predict **True** or **False**.

2. Identify missing axioms (if True)
Find H such that $\mathcal{O}' \cup H \models \alpha$.

Example: every pet cat is an animal

Query α
`PetCat` $\sqsubseteq$ `Animal` ?

incomplete ontology $\mathcal{O}'_1$:



✗ **OWL reasoner** — **No**
- α is not entailed by $\mathcal{O}'_1$

? **TBox abduction** — *predefined scope*
- finds `PetCat` $\sqsubseteq$ `Cat` only if in scope
- pure logic: no textual information
- assumes α holds: never **No**

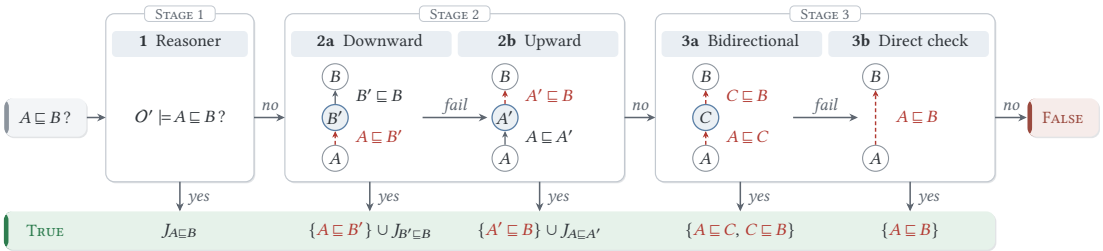
✓ **NeurOWL** — **Yes**
- completes $\mathcal{O}'_1$ with `PetCat` $\sqsubseteq$ `Cat`
- no predefined scope needed
- integrates textual information
- implausible queries: **No**

Yang et al., NeurOWL (preprint, revised version) [91], Fig. 1 and problem formulation

NeurOWL[91]: overall method

preprint

Three-stage reasoning over an incomplete ontology



1. Existing proof

The reasoner checks entailment and returns a justification.

2. One missing axiom

Test downward and upward bridges using known subsumptions.

3. Bridge or direct check

Test two bridge links, then the target directly; return False if all fail.

Yang et al., NeurOWL (preprint, revised version) [91], Fig. 2 and §4 (overall framework)

3

Summary

Three takeaways
Future Work

Summary: three takeaways

1. **LMs can prove, but struggle with complex cases**
   - GPT-o4-mini: 80% whole-proof accuracy on standard SNOMED CT cases.
   - On the hard cases, whole-proof accuracy falls to 32%.
2. **Noise limitations**
   - Distractors hurt extraction, simplification, and derivation.
   - Natural-language wording has a limited effect in these tests $\Rightarrow$ logical complexity is the main bottleneck.
3. **Textual information supports abductive reasoning**
   - NeurOWL uses textual information to identify and assess candidate axioms.

Yang et al., WWW 2026 [60], Tables 2–3 and §6.2; Yang et al., NeurOWL [91], §4

Future Work

1. **Scale LM proof generation to large, complex ontologies**
   - Design retrieval structures that supply proof-relevant ontology axioms.
   - Use retrieval and filtering to select reliable, relevant premises.
2. **Analyze proofs from informal input**
   - Analyze general reasoning processes expressed entirely in natural language.
3. **Extend abduction to more complex missing knowledge**
   - Improve performance on structurally complex missing axioms.
   - Integrate abduction with ontology repair and ontology construction.

Research directions building on LLM4Proof [60], §7, and NeurOWL [91]

References

[60] Yang, Chen, Sattler. Large Language Model for OWL Proofs (LLM4Proof). WWW 2026. github.com/HuiYang1997/LLMOwlR.

[65] Saparov & He. Language Models Are Greedy Reasoners: A Systematic Formal Analysis of Chain-of-Thought (PrOntoQA). ICLR 2023.

[81] Zhang, Lai, Meng, Bos. OntoURL: A Benchmark for Evaluating LLMs on Symbolic Ontological Understanding, Reasoning and Learning. J. Web Semantics 91, 2026.

[91] Preprint: Yang, Chen, Song, Schmidt, Zhang. NeurOWL: An LLM-Based Neural-Symbolic Framework for Incomplete OWL Ontology Reasoning. arXiv 2607.15776, 2026.