NeuroSym4MLLM - Atelier EGC 2026
Anglet, 27 janvier 2026

Ontology Guided Large Language Model Pipeline for Structured Information Extraction from Battery Cell Datasheets

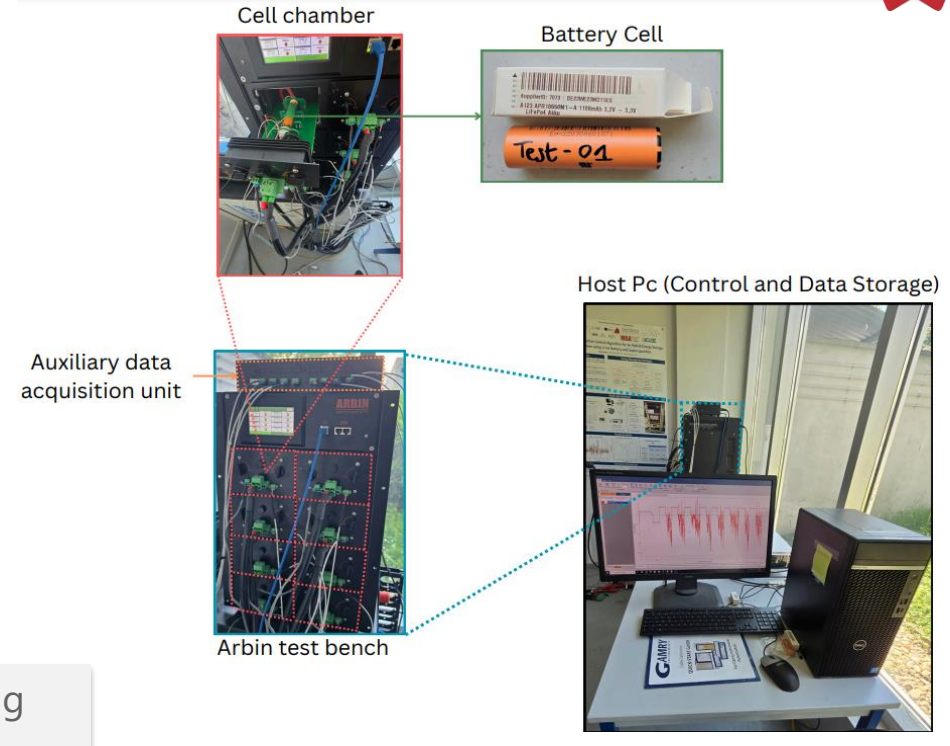
Slimane Arbaoui, Ali Ayadi, Ahmed Samet, Tedjani Mesbahi and Romuald Boné

Université de Strasbourg, INSA Strasbourg, ICube Laboratory UMR 7357

Context & Motivation

The Challenge

- Electric vehicles rely on battery cells as core energy-storage units
- Critical parameters needed: charging current, temperature range, voltage limits
- Datasheets often in PDF format, unstructured, multilingual (especially Chinese manufacturers)
- Information inconsistent, incomplete, or ambiguous across manufacturers



Key Problem: Manual extraction is time-consuming and error-prone. Automated extraction is essential for accurate battery modeling and testing.

Research Objectives

1
Automated Extraction Pipeline

2
Reduce LLM Hallucinations

3
Multi-language Support

Our Contributions

- **Dual-LLM pipeline** combining Chain-of-Thought (CoT) and few-shot prompting
- **Ontology-driven approach** to ensure structured, validated output
- **Scalable solution** for heterogeneous, multi-language datasheets

INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES STRASBOURG
energetic

Related Work: LLMs for Knowledge Extraction

Successes

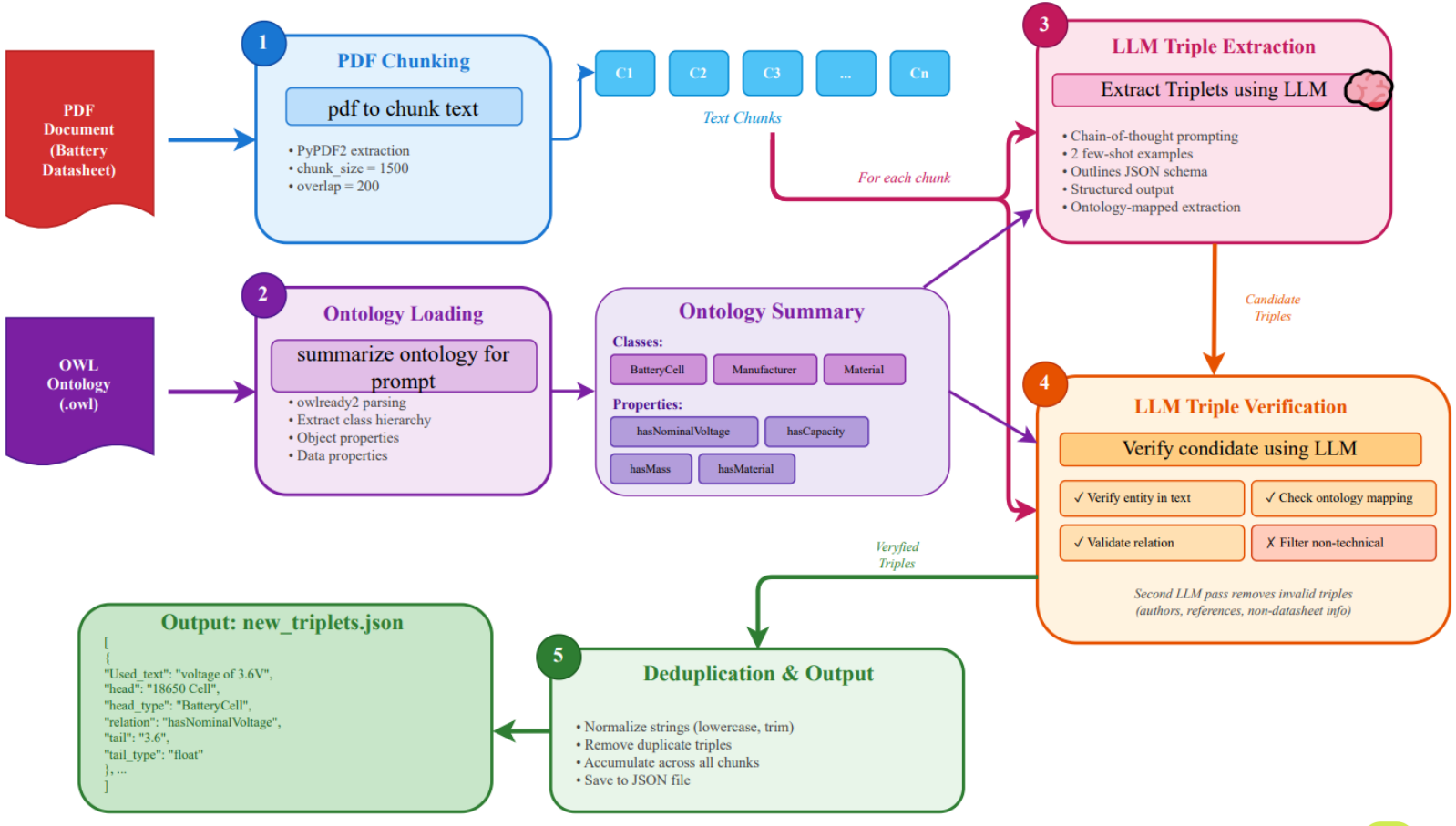
- LLMs outperform traditional NLP for entity/relation extraction
- Zero-shot KG construction from scientific papers
- Domain-specific extraction in materials science

Challenges

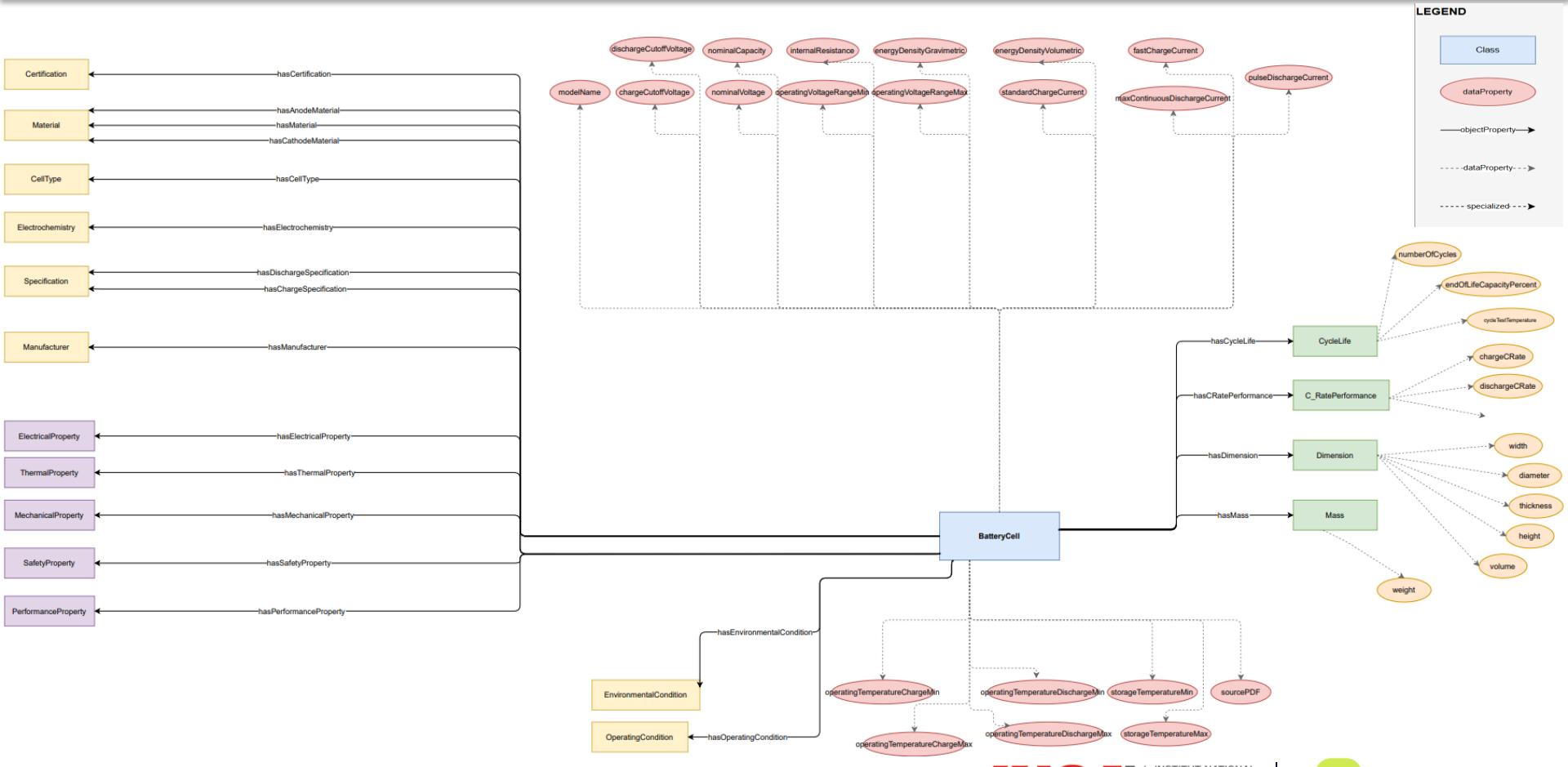
- **Hallucinations:** plausible but incorrect outputs
- Inconsistent entity resolution
- Semantic duplication
- Technical terminology handling

Our Solution: Hybrid approach combining LLM extraction + ontology-guided verification

Methodology: Three-Component Pipeline



Component 1: Battery Cell Ontology



INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES STRASBOURG

energetic

Component 1: Battery Cell Ontology

Q#	Competency Question	Answerable?	Explanation
1	What is the manufacturer of a specific battery cell?	Yes	Use `hasManufacturer` linking `BatteryCell` to `Manufacturer`.
2	What is the model name of a battery cell?	Yes	Use `hasModelName` of `BatteryCell`.
3	What is the nominal voltage and capacity of a battery cell?	Yes	Use `hasNominalVoltage` and `hasNominalCapacity`.
4	What is the internal resistance of a battery cell?	Yes	Use `hasInternalResistance`.
5	What is the charge and discharge cutoff voltage?	Yes	Use `hasChargeCutoffVoltage` and `hasDischargeCutoffVoltage`.
6	What is the operating voltage range?	Yes	Use `hasOperatingVoltageRangeMin` and `hasOperatingVoltageRangeMax`.
7	What is the standard and fast charge current?	Yes	Use `hasStandardChargeCurrent` and `hasFastChargeCurrent`.
8	What is the max continuous and pulse discharge current?	Yes	Use `hasMaxContinuous DischargeCurrent` and `hasPulseDischargeCurrent`.
9	What are the operating temperatures for charge and discharge?	Yes	Use `hasOperatingTemperature ChargeMin/Max` and `hasOperatingTemperature DischargeMin/Max`.
10	What are the storage temperature limits?	Yes	Use `hasStorageTemperatureMin` and `hasStorageTemperatureMax`.
11	What are the physical dimensions of a battery cell?	Yes	Use `Dimension` class with `hasDiameter`, `hasWidth`, `hasHeight`, `hasThickness`, `hasLength`, `hasVolume`.
12	What is the weight of a battery cell?	Yes	Use `Mass` class with `hasWeight`.

Component 2: Extraction LLM with CoT & Few-Shot

Prompting Strategy

- **Chain-of-Thought (CoT):** Step-by-step reasoning guidance
- **Few-shot examples:** Sample triples showing desired output format
- **Ontology summary:** Classes and properties in prompt (not full OWL)
- **Rules:** Process sentence-by-sentence, normalize entities, translate to English

Why this approach?
CoT + few-shot significantly improves extraction without model fine-tuning, while keeping token usage manageable

Component 2: Extraction LLM with CoT & Few-Shot

You are a battery datasheet information extraction model. Your task is to extract **only technical datasheet information** such as: - Electrical specifications: nominal voltage, capacity, charge/discharge limits, internal resistance - Mechanical specifications: dimensions, mass - Materials & components: electrodes, separators, electrolytes - Cell type / chemistry

RULES:

1. Process the text sentence by sentence.
2. Map each entity to a class from the ontology dictionary.
3. Use only valid relations (object/data properties) from the ontology.
4. Normalize entity names to full canonical identifiers.
5. Return ONLY triples aligned with the ontology and found in the text. Output format: JSON array of objects with keys: "Used_text", "head", "head_type", "relation", "tail", "tail_type"

Ontology dictionary (classes and relations):
{ontology_summary}
Few-shot examples:
{examples_str}
Text to analyze:
{chunk_text}

Extraction Example

Input Text Chunk

"The 18650 lithium-ion cell has a nominal voltage of 3.6V and a nominal capacity of 3000mAh."

Extracted Triple (JSON format)


```
{ "Used_text": "The 18650 lithium-ion cell has...",
"head": "18650 LithiumIonCell",
"head_type": "BatteryCell",
 "relation": "hasNominalVoltage",
"tail": "3.6",
 "tail_type": "float" }
```

Component 3: Validation LLM as Judge

Validation Criteria

- **Ontology compliance:** head_type, tail_type, relation must exist in ontology
- **Text grounding:** Information must be present in source chunk
- **Format validation:** Correct JSON structure with required keys
- **Content filtering:** Only technical datasheet information retained

Result: Reduces hallucinations, and non-datasheet content (authors, references, etc.)

Component 3: Verification LLM (Judge)

You are an expert validator for battery datasheet information. Your task is to strictly validate candidate triples.

RULES:

1. Keep only triples that describe **technical datasheet information**: - Electrical properties: voltage, capacity, charge/discharge limits, internal resistance - Mechanical properties: dimensions, mass - Materials & components: electrodes, electrolytes, separators - Cell type / chemistry
2. Verify that each entity exists in the text chunk.
3. Verify ontology alignment: head_type, tail_type, relation must exist in ontology.
4. Ignore all non-battery content (authors, references, journals, etc.)

Output ONLY a JSON array of objects with keys: "Used_text", "head", "head_type", "relation", "tail", "tail_type"

Candidate triples:
{candidates}
Text chunk:
{chunk_text}
Ontology dictionary:
{ontology_summary}
Return ONLY the JSON array of validated datasheet triples.

Experiments: Setup & Models

Tested Models (5 Lightweight LLMs)

- Qwen/Qwen2-7B-Instruct
- mistralai/Mistral-7B-Instruct-v0.3
- meta-llama/Llama-3.1-8B-Instruct
- Qwen/Qwen3-8B
- openai/gpt-oss-20b

Evaluation

- 5 PDFs from major manufacturers
- Same LLM for extraction & verification(Manual validation of extracted triples
- Strict criteria: no incomplete sentences, errors, or duplicates

Results: Extracted vs Validated Triples

Model	P1	P2	P3	P4	P5
Qwen2-7B	34 / **19**	117 / **30**	70 / **13**	62 / **25**	70 / **17**
Mistral-7B	7 / **1**	41 / **6**	26 / **1**	13 / **0**	18 / **0**
Llama-3.1-8B	26 / **5**	120 / **20**	76 / **16**	68 / **13**	66 / **13**
Qwen3-8B	56 / **29**	157 / **20**	81 / **23**	83 / **31**	69 / **23**
gpt-oss-20b	12 / **7**	46 / **7**	21 / **5**	19 / **2**	20 / **3**

Format: Total Extracted / **Validated**

Key Findings

Performance Insights

- **Best performers:** Qwen3-8B and Qwen2-7B-Instruct (most validated triples)
- **Significant variance:** Model choice critically impacts extraction quality
- **Validation effectiveness:** Filters out ~50-70% of raw extractions
- **Recent models win:** Newer architectures show better structural consistency

Conclusion & Future Directions

✓ Achievements

- Effective ontology-driven pipeline
- Reduced hallucinations via dual-LLM verification
- Multi-language support demonstrated
- No fine-tuning required

→ Future Work

- Automated evaluation metrics
- Extended ontology coverage
- Integration with knowledge graphs

Code available: github.com/BatIE-OGLLM

Funding: ANR-22-CE92-0007-02 & Horizon Europe GAP-101103667

INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES STRASBOURG

energetic