Integrating Large Language Models and Knowledge Graphs for Extraction and Validation of Textual Test Data

Antonio De Santis[1], Marco Balduini[2,3], Federico De Santis[3], Andrea Proia[4], Arsenio Leo[4], Marco Brambilla[1], and Emanuele Della Valle[1]

[1] Politecnico di Milano, DEIB, I-20133 Milano, Italy
{firstname.lastname}@polimi.it
[2] Quantia Consulting, Milano, Italy
marco.balduini@quantiaconsulting.com
[3] motus ml, Milano, Italy
{firstname.lastname}@motusml.com
[4] Thales Alenia Space, Roma, Italy
{firstname.lastname}@thalesaleniaspace.com

Abstract. Aerospace manufacturing companies, such as Thales Alenia Space, design, develop, integrate, verify, and validate products characterized by high complexity and low volume. They carefully document all phases for each product but analyses across products are challenging due to the heterogeneity and unstructured nature of the data in documents. In this paper, we propose a hybrid methodology that leverages Knowledge Graphs (KGs) in conjunction with Large Language Models (LLMs) to extract and validate data contained in these documents. We consider a case study focused on test data related to electronic boards for satellites. To do so, we extend the Semantic Sensor Network ontology. We store the metadata of the reports in a KG, while the actual test results are stored in parquet accessible via a Virtual Knowledge Graph. The validation process is managed using an LLM-based approach. We also conduct a benchmarking study to evaluate the performance of state-of-the-art LLMs in executing this task. Finally, we analyze the costs and benefits of automating preexisting processes of manual data extraction and validation for subsequent cross-report analyses.

Keywords: Knowledge Graphs · Large Language Models · Data Extraction · Space Industry

1 Introduction

Context. Companies in the aerospace industry produce complex products in low volumes. As a result, most of the data that can boost analytics is hidden within documents, making its extraction challenging. The experience presented in this article focuses on Test Data related to electronic boards used in Thales Alenia Space's satellite systems. The production of these electronic boards is a

critical aspect of space technology [24]. These boards are manufactured in limited quantities, with a satellite containing between 10 to 20 such boards. Moreover, these components must be extremely reliable and are subject to rigorous testing protocols due to the hostile conditions of space missions [11]. Given the near impossibility of conducting repairs once satellites are in space, production errors could potentially lead to the failure of an entire mission, which would result in significant financial losses and wasted resources. In this scenario, data analytics can play a crucial role, providing timely insights and enabling immediate actions based on the data's flow and characteristics. For example, the analysis of historical production data could reveal trends that can predict the likelihood of future components failing the quality tests. Such insights can guide production decisions, minimizing waste and resulting in significant cost savings.

Problem Statement. The effectiveness of these data-driven approaches relies on the quality and organization of the data [21]. Each electronic board is meticulously crafted and tested before receiving approval. However, the testing procedures and the generation of Test Reports are manually executed by human operators across multiple isolated documents (primarily in .docx and .pdf format). This leads to data that is highly fragmented, heterogeneous, unstructured, and prone to errors and inconsistencies. Such a scenario poses a significant challenge, as it can jeopardize data analysis efforts. Considering this, the focus of our case study is automating the extraction, validation, and integration of Test Data. Given the high level of data heterogeneity, the process of validation is particularly challenging because a standard approach based on regular expressions would be impractical.

Proposed Solution. To address the aforementioned challenges, we propose a hybrid approach that utilizes Large Language Models (LLMs) in combination with Semantic Web technologies. To provide semantic knowledge to the system and manage structural heterogeneity, we create an ontology to capture the semantics of the data. This ontology extends the *Semantic Sensor Network* (SSN) [7] ontology, a well-established ontology for representing sensor data. We then proceed with extracting the data from Test Report documents and storing it in tabular format. The extracted data must undergo an automatic validation process (i.e., checking for inconsistencies in test results). For this task, we exploit the implicit knowledge of LLMs. These models have demonstrated their capability to process data despite structural and syntactic heterogeneity. Moreover, in contrast with approaches based on regular expressions, LLMs have the advantage of being able to scale effectively with an increasing variety and complexity of the data. The validated data is integrated into a data storage system, ensuring a structured and organized data repository. To facilitate direct data access, we then create mappings between the data storage and the ontology, allowing our system to understand the relationships and connections among data points. This knowledge is stored in a Virtual Knowledge Graph (VKG) [39], also known in the literature as Ontology-Based Data Access (OBDA) [38], and is accessed using SPARQL queries, which are automatically translated into SQL language.

Structure of the Work. The paper is structured as follows. Section 2 presents a review of related work. Section 3 describes the case study in detail and Section 4 explains the rationale behind using KGs and LLMs. Section 5 presents our methodology, whose implementation and evaluation are detailed in Section 6. Section 7 discusses the uptake of our work and the lessons learned, while Section 8 concludes the paper, providing directions for future work.

2   Related Work

Industrial deployment of VKGs. Semantic Web technologies have been successfully applied in several industrial contexts [32, 12, 5] as they simplify data access by providing an abstraction layer (i.e., an ontology) that integrates data from semantically and physically different sources. Siemens uses an OBDA for managing the temperature data of trains and turbines and developed a semantic rule-based diagnostic system [18, 17, 4]. Statoil has implemented an OBDA using the *Ontop* [31] framework for integrating multiple data sources [16]. This system has enhanced the efficiency of data collection for geologists in the field of oil and gas exploration and production. Similarly, Ontop was used to realize a semantic information model for managing machine data [28]. Moreover, Ford Motor Company also stores knowledge about manufacturing processes in an ontology [33]. This allows their internally developed AI system to handle the planning of vehicle assembly processes. They have also explored the use of federated ontologies to identify potential risks in the supply chain [26, 19]. Bosch also has utilized ontology-based approaches for data access. They applied knowledge graph embedding [34] and ontology reshaping [41] for automatic knowledge graph (KG) construction in a case related to welding quality monitoring.

LLMs for Data Management. In recent years, the field of language models has experienced substantial progress due to the introduction of LLMs such as GPT-3.5 [40] and GPT-4 [25], developed by OpenAI, Meta's Llama 1 [35] and Llama 2 [36], Claude 3 [2] from Anthropic, Google's Gemini [13], and Mixtral [14], from Mistral AI. These models have been utilized in a variety of data management tasks [42, 43] due to their ability to extract knowledge from unstructured data sources [1] and to understand the data without the need for explicit modeling [10]. From a data validation perspective, LLMs have demonstrated close to human-level capabilities in detecting inconsistencies in text summaries [20]. In the context of the Semantic Web, LLMs can also be used to automate KG completion and construction [27]. For instance, GPT-4 was used for automatic ontology and KG construction for large amounts of unstructured sustainability-related data [37]. Moreover, LLMs have been effectively utilized to assist with data preparation tasks required before performing business analytics [23]. More specifically, GPT-4 was used to translate product names, assign product categories, classify customer sentiment, and extract repair requests and their causes from customer service logs. Regarding real industry scenarios, there is currently limited evidence, to our knowledge, of LLMs being utilized in conjunction with semantic technologies for data validation in large manufacturing companies.

Test_Report1_Board1			Test_Report2_Board1			Test_Report1_Board2			Test_Report2_Board2			Test_Report1_Board3		
9	TEST SEQUENCE	65	9	TEST SEQUENCE	65	9	TEST SEQUENCE	70	9	TEST SEQUENCE	65	9	TEST SEQUENCE	58
10	***	67	10	***	67	10	***	72	10	***	67	10	***	59
11	***	69	11	***	69	11	***	74	11	***	69	11	***	61
12	ISOLATION	70	12	ISOLATION	70	12	ISOLATION	75	12	ISOLATION	70	12	ISOLATION	62
12.1	External Isolation Measure	70	12.1	External Isolation Measure	70	12.1	External Isolation Measure	75	12.1	External Isolation Measure	70	12.1	External Isolation Measure	62
12.2	Internal Isolation Measure	72	12.2	Internal Isolation Measure	72	12.2	Internal Isolation Measure	77	12.2	Internal Isolation Measure	72	12.2	Internal Isolation Measure	63
13	P.O.L. Voltage Verification	73	13	P.O.L. Voltage Verification	73	13	P.O.L. Voltage Verification	78	13	P.O.L. Voltage Verification	73	13	P.O.L. Voltage Verification	65
14	Preliminary Power Consumpt	75	14	Preliminary Power Consumpt	75	14	Preliminary Power Consumpt	80	14	Preliminary Power Consum	75	14	Preliminary Power Consumpt	67
14.1	***	75	14.1	***	75	14.1	***	80	14.1	***	75	14.1	***	67
15	***	76	15	***	76	15	***	81	15	***	76	15	***	68
15.1	***	78	15.1	***	78	15.1	***	82	15.1	***	77	15.1	***	69
15.1.1	***	78	15.1.1	***	78	15.1.1	***	82	15.1.1	***	77	15.1.1	***	69
15.1.2	***	79	15.1.2	***	79	15.1.2	***	83	15.1.2	***	78	15.1.2	***	70
												15.1.3	***	71
						15.1.3	***	84	15.1.3	***	79	15.1.4	***	71
15.1.3	***	80	15.1.3	***	80									
15.1.3.1	***	80	15.1.3.1	***	80									
15.1.3.2	***	81	15.1.3.2	***	81									
15.1.3.3	***	81	15.1.3.3	***	81									
15.1.3.4	***	82	15.1.3.4	***	82									
15.2	***	83	15.2	***	83	15.2	***	85	15.2	***	80	15.2	***	72
						15.2.1	***	85	15.2.1	***	80	15.2.1	***	72
						15.2.1.1	***	85	15.2.1.1	***	80	15.2.1	***	72
												15.2.1	***	73
						15.2.1.2	***	86	15.2.1.2	***	81	15.2.1	***	73
15.2.1	***	83	15.2.1	***	83	15.2.1.3	***	86	15.2.1.3	***	81	15.2.1	***	74
15.2.2	***	85	15.2.2	***	85	15.2.2	***	88	15.2.2	***	83	15.2.2	***	76
												15.2.2	***	76
												15.2.2	***	77

Fig. 1. A portion of a color-coded spreadsheet that visually represents the heterogeneity within Test Reports, which typically contain around 23 sections. Green denotes uniform sections, while yellow represents variable ones. White cells indicate the absence of a section. Titles are intentionally obscured to protect confidential information.

3   Case Study: Testing of Electronic Boards

In this section, we discuss our case study in greater detail and describe the structure and characteristics of Thales Alenia Space's Test Reports.

Electronic Boards Test Data. Our case study involves Test Data for electronic boards, primarily Printed Circuit Boards (PCBs) used in satellite systems. Testing these products is a critical process in the space industry, ensuring that all technological processes meet specific mission requirements and comply with standards established by the European Space Agency (ESA) and the European Cooperation for Space Standardization (ECSS). The tests involve measuring parameters such as voltage, resistance, or power, and comparing the results to a predefined expected range, which represents the acceptable limits within which the parameter should fall for the PCB to operate correctly. Test engineers conduct these tests, which are documented in Test Reports. These documents, which are primarily in .docx and .pdf format, are manually filled by the engineers and exhibit a high degree of heterogeneity. In Figure 1, we show a color-coded spreadsheet to illustrate the heterogeneity within these documents. The actual test results in the reports are organized within manually filled tables. The "acceptance limits" column is pre-filled and the engineers have to fill in the measured value and a "successful" column based on the test outcome. In this study, we consider Point-of-Load (POL) Voltage Verification, Preliminary Power Consumption, and Isolation (both external and internal) as representative types of tests. Figure 2 provides an example of tables for these types of tests, emphasizing the unstructured and heterogeneous nature of the data which manifests in several ways:

Syntactic Heterogeneity. This is seen in the different formatting of the data. The range of acceptance limits is represented in various ways. For instance, "[3.198,

PRELIMINARY POWER CONSUMPTION

POWER LINE	CURRENT CONSUMPTION T_{AMB}	Acceptance limits	
+3.35V	1.090A	< 1.400A	
+6.35V	0.022A	< 30mA	

P.O.L. VOLTAGE VERIFICATION

V CORES	Voltage Measurements [V] T_{AMB}	Acceptance limits	Successful
Core1	1.097	[1.076, 1.224] V	OK
Core2	2.508	[2.470, 2.530] V	OK
Core3	1.801	[1.864, 1.936] V	
Core4	2.206	[2.000, 2.650] V	OK
Core5	3.968	[3.168, 4.432] V	OK
Core6	3.374	[3.198, 3.532] V	OK

External Isolation Measurement

Check between points		Measured Values	Expected Values	Successful
+3.3V	GND	787 Ω	> 100 Ω	OK
+14.5V	GND	2 MΩ	> 100 KΩ	OK
+24V	GND_HPC	11 KΩ	> 5 KΩ	OK
+2.3V	+5V	33 K	> 20 KΩ	OK
+2.3V	+13.5V	2 MΩ	> 100 KΩ	OK
+9.5V	+29V	2 MΩ	> 100 KΩ	OK
GND	GND_HPC	1.38 MΩ	1.1M – 1.9MΩ	OK

Internal Isolation Measurement

Check between points		Measured Values	Expected Values	Successful
JP1_pin1	GND	97 Ω	> 100 Ω	-
JP1_pin2	GND	460 Ω	> 100 Ω	OK
JP2_pin1	GND	885 Ω	> 100 Ω	OK
JP2_pin2	GND	620 Ω	> 100 Ω	OK
JP4_pin1	GND	2 KΩ	> 100 Ω	OK
JP5_pin2	GND	7.6 KΩ	> 100 Ω	OK

Fig. 2. Examples of test results tables that illustrate the challenges of syntactic (shown in green), structural (shown in yellow), and semantic (shown in purple) heterogeneity.

3.532] V" and "1.1M - 1.9MΩ" both indicate a range of acceptance. In some cases, the measured value and the acceptance limits are indicated with different units of measure. Additionally, in the "successful" column, the absence of a value or the presence of a "-" both indicate a lack of success.

Structural Heterogeneity. This is evident in the inconsistent organization and naming of the tables. For instance, some tables have a single "successful" cell in a different part of the document and therefore lack a dedicated "successful" column. Furthermore, a column labeled "Acceptance limits" in one table might be labeled as "Expected Values" in another. The unit of measure can be included in the table title as well as written with the values or even absent. Another form of structural heterogeneity can be observed in the use of row span, which is used to indicate that the same value applies to multiple rows.

Semantic Heterogeneity. There is an implicit hierarchical structure within the reports as there are various representations for the concept of a "Test", such as "Internal Isolation", "External Isolation" or "POL Voltage". These test types share many properties, but they are categorized separately due to their specific aspects. Similarly, Internal and External Isolation fall under the category of Isolation tests, each possessing properties specific to Isolation testing. Despite this, they are represented differently, introducing a semantic heterogeneity. This leads to the requirement of modeling what is a "Test" or an "Isolation test".

The preexisting manual approach (see Figure 3) for data extraction and validation is costly, time-consuming, and allows for limited cross-report analyses, but automating these processes isn't straightforward. Although a human operator can intuitively understand that, for example, "Acceptance limits" has the same meaning as "Expected values", this poses a challenge for an automated system.
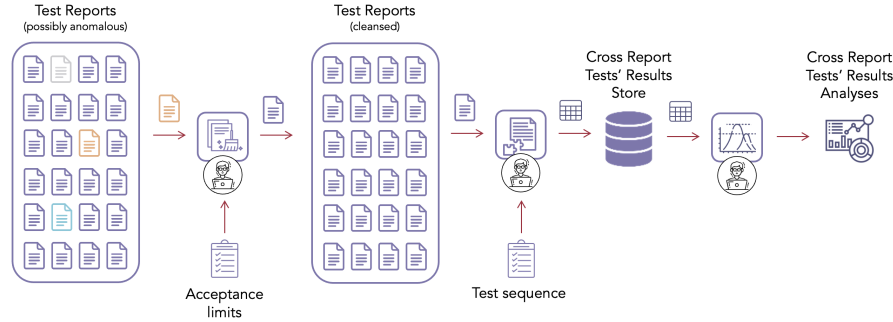
Fig. 3. The preexisting manual data processing workflow in which potentially anomalous reports are subjected to manual extraction, cleaning, and validation.

Data Obfuscation. Data is not disclosed in its original form to protect Thales Alenia Space's privacy. We added noise to the values, ensuring the structure and syntax remained intact without disclosing any confidential information.

4 Motivation

In this section, we aim to clarify our motivation by addressing two key questions: (1) Why do we need KGs? and (2) Why integrate them with LLMs?

Motivation for Knowledge Graphs. The motivation for choosing KGs and OBDA systems lies in their ability to handle heterogeneous and physically distributed information, a common challenge in knowledge-intensive industries such as aerospace. KGs effectively accommodate the high diversity and low volume of data in the space industry, which produces hundreds of PCB families (with similar but not identical designs) but only a few dozen PCBs. The industry also deals with a diverse array of tests due to the intricate nature of PCBs, which include passive and active electrical components, as well as digital electronics like RAM, CPUs, and FPGAs. Leveraging and extending resources such as the SSN ontology can facilitate the modeling process in this case. Furthermore, a graph-based representation allows for a more explicit data repository, reducing the reliance on tacit knowledge held by domain experts. This is crucial in aerospace where semantic coherence is key for managing complex systems such as satellites.

Integrating LLMs with KGs. Consider the detailed RDF representation in Listing 1.1 that includes the QUDT (Quantities, Units, Dimensions, and Types) [9] ontology for the units of measurement. Annotating data in this way would require a large amount of manual work at the level of the template of the Test Report. This can be challenging and time-consuming when dealing with complex and diverse data. Moreover, the complexity grows with the number of different templates of Test Reports the company introduces (i.e., one per PCB family). See once again Figure 1 to feel the degree of heterogeneity at the level

of the sections of the reports. However, LLM's ability in natural language understanding can determine whether a measured value falls within an expected range, even if the syntax changes or the units of measurement differ. Therefore, it can assist in error detection and simplify the modeling process. This leads to a lightweight annotation of the data (see Listing 1.2) using the Unified Code for Units of Measure (UCUM) [22], allowing data engineers to focus on the conceptual model and semantic meaning of the data, without having to account for every minor syntactic heterogeneity.

```
<http://tasi.com/pol#TASI-1234-Core1> a sosa:Observation ;
    rdfs:label "TASI-1234-Core1" ;
    sosa:observedProperty tasi:POLVoltage ;
    sosa:hasResult [
      a qudt-1-1:QuantityValue ;
      qudt-1-1:numericValue "1.097"^^xsd:double ;
      qudt-1-1:unit qudt-unit-1-1:Volt
  ] ;
  tasi:hasAcceptanceLimits [
      a tasi:Range ;
      tasi:lowerLimit [
         a qudt-1-1:QuantityValue ;
         qudt-1-1:numericValue "1.076"^^xsd:double ;
         qudt-1-1:unit qudt-unit-1-1:Volt
      ] ;
      tasi:upperLimit [
         a qudt-1-1:QuantityValue ;
         qudt-1-1:numericValue "1.224"^^xsd:double ;
         qudt-1-1:unit qudt-unit-1-1:Volt
      ] ;
  tasi:hasTestResult "OK" ;
  tasi:reportedIn "TASI-1234" ;
  tasi:testReportDate "2023-06-15"^^xsd:dateTime .
```

Listing 1.1. Detailed representation that a machine can understand without LLMs.

```
<http://tasi.com/pol#TASI-1234-Core1> a sosa:Observation ;
    rdfs:label "TASI-1234-Core1" ;
    sosa:observedProperty tasi:POLVoltage ;
    sosa:hasSimpleResult "1.097 V"^^cdt:ucum ;
    tasi:hasAcceptanceLimits "[1.076, 1.224] V" ;
    tasi:hasTestResult "OK" ;
    tasi:reportedIn "TASI-1234" ;
    tasi:testReportDate "2023-06-15"^^xsd:dateTime .
```

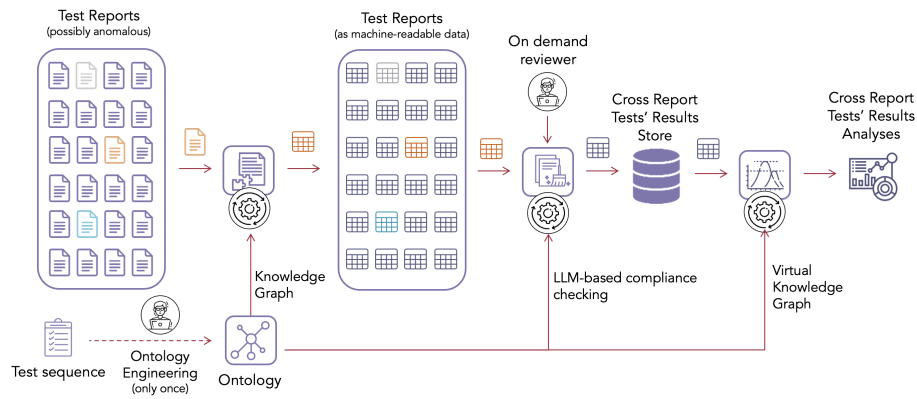**Listing 1.2.** Lightweight representation that can be understood by LLMs.

Fig. 4. A flowchart representation of the proposed methodology. The process begins with the input of a set of potentially anomalous Test Reports, from which the data is extracted and transformed into a machine-readable format. The documents' metadata is integrated into a KG, while the test results undergo LLM-based compliance checking and anomalies are handled by an on-demand reviewer. The validated data is accessed through a VKG, enabling access to heterogeneous data and facilitating cross-report analyses. The whole process is guided by a one-time ontology engineering process.

5 Methodology

In this section, we describe the methodology of our approach for extraction, validation, and integration of Test Data from unstructured Test Reports. As depicted in Figure 4, the process is divided into several phases. The validation process is managed using an LLM-based approach. On the other hand, data integration is accomplished through KGs, enabling access to heterogeneous data. More specifically, the process is structured on three levels:

- *Data Extraction:* Test Reports' metadata and the test types they contain are extracted and stored in a KG using an ontology.
- *LLM-Based Compliance Checking:* LLMs are used to validate that test results are consistent with their respective acceptance limits.
- *Ontology-Based Data Access:* A VKG is used to mediate the actual access to the test results.

Data Extraction. The first step in our process is to extract the textual data within the Test Reports and transform it into a machine-readable format. This transformation is facilitated by a one-time ontology engineering process that defines the concepts, categories, and relationships embedded within the data. A KG is used for this purpose. The ontology used in this KG is an extension of the SSN ontology (see Figure 5) and maps the information related to Test Reports and observable properties found within these reports, which refer to the property being tested (i.e., the test type) and the related test table structure description in terms
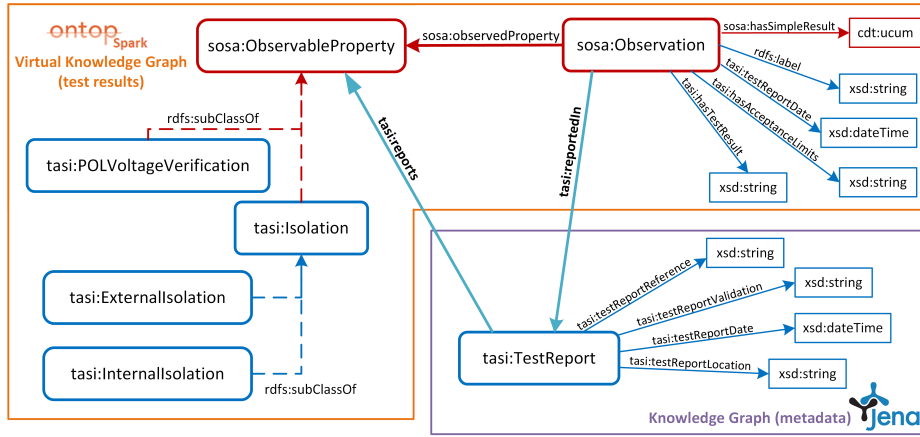
Fig. 5. The figure provides a visual representation of our ontology. This ontology is an extension of the well-established SSN ontology, which is denoted by the `sosa` prefix and the red color. Our additions include new classes and properties, which are identified by the `tasi` prefix and the blue color. The ontology's components used for modeling Test report metadata are enclosed within a purple rectangle, and this metadata is stored within a KG. The modeling of test results, represented by an orange rectangle, is stored in a structured data repository and made accessible via a VKG.

of its columns. All tests are of type `sosa:ObservableProperty` with their respective hierarchy. For instance, `<POLVoltage>` and `<Isolation>` are defined as `sosa:ObservableProperty`. `<InternalIsolation>` and `<ExternalIsolation>` are defined as sub-classes of `<Isolation>`. The RDF fragment provided in Listing 2 is an example of how a Test Report is modeled. An additional property `tasi:reports` has been added due to the absence of a Test Report concept in the SSN ontology. A Test Report is defined as `<TestReport>` and is associated with the observable properties such as `<InternalIsolation>`, `<ExternalIsolation>` and `<POLVoltage>`. The metadata of the report is modeled using additional properties such as `testReportDate`, `testReportLocation`, `testReportName`, `testReportReference` and `testReportValidation`. The latter is used to indicate whether the whole Test Report is valid.

Using the ontology definition as a basis, we can streamline the extraction process. The procedure begins with parsing the Test Reports to identify relevant sections. These reports are then extracted along with their observable properties, such as POL Voltage, using the KG test table structure definition to automatically identify the purpose of each column (i.e., for the `POLVoltage` table, `Voltage Measurements [V]` contains the test data entry, while `Acceptance Limits` contains the entry validation range). Subsequently, this data is transformed into RDF triples and stored in the KG. The creation of these RDF triples is guided by the ontology, ensuring that the resulting data is both structured and machine-readable. The actual observations, which correspond to the rows in the tables, are extracted and temporarily stored in a data repository for subsequent validation.

```
@prefix tasi:    <http://www.semanticweb.org/ontologies/tasi#> .

tasi:POLVoltage a sosa:ObservableProperty ;
  rdfs:label "P.O.L. Voltage"@en .
  tasi:obsPropertyAccLimLocation  "VALIDATED/ist/TASI-1234-ist_pol.csv" ;
  tasi:obsPropertyResultsLocation "/VALIDATED/pol/pol.parquet" .
tasi:Isolation a sosa:ObservableProperty ;
  rdfs:label "Isolation"@en .
tasi:InternalIsolation rdfs:subClassOf tasi:Isolation ;
  rdfs:label "Internal Isolation"@en .
  tasi:obsPropertyAccLimLocation  "VALIDATED/ist/TASI-1234-ist_al.csv" ;
  tasi:obsPropertyResultsLocation "VALIDATED/ist/ist.parquet" .
tasi:ExternalIsolation rdfs:subClassOf tasi:Isolation ;
  rdfs:label "External Isolation"@en .

<http://tasi.com#TASI-1234> a tasi:TestReport ;
  tasi:reports tasi:POLVoltage, tasi:InternalIsolation ;
  tasi:testReportDate "2023-06-15"^^xsd:dateTime ;
  tasi:testReportName "test_report_xy" ;
  tasi:testReportReference "TASI-1234" ;
  tasi:testReportValidation "OK" ;
  tasi:testReportLocation "path/to/test_report_file.docx" .
```

Listing 2. An example of a Test Report modeled using an extension of the SSN ontology that contains valid data for the `POLVoltage` and `InternalIsolation` tests.

LLM-Based Compliance Checking. The primary challenge in managing Test Data lies in the expensive and time-consuming task of compliance checking. This process is difficult to automate algorithmically due to the high heterogeneity in observed values and the wide variety of formats used for the acceptance limits. However, compliance checking can be automated using LLMs, as these models are capable of handling data with syntactic and structural heterogeneity. This ability makes the compliance checking process applicable across a broad spectrum of testing scenarios. Consequently, data engineers can focus only on a small subset of tests that the LLM identifies as anomalous. The validation process is conducted row by row, rather than for the whole table at once, to prevent disclosing confidential information. For each test result, we prompt the LLM to determine whether the measured value is within the acceptance range. The LLM's response is then compared with the "successful" value. If there is a mismatch between these two values, the test is classified as anomalous. A test is considered valid if the measured value is within the predefined acceptance limits and the "successful" column reads "OK", or if the value is outside the range and the "successful" column does not read "OK".

The prompt strategy chosen is the *Zero-shot* [29] (i.e., direct prompting without any examples) using a task description instead of a role-oriented approach. For data validation tasks, this strategy was shown to be superior, especially for

bigger models [20]. This is consistent with previous findings showing that zero-shot prompts are best when the task involves utilizing pre-existing knowledge embedded within the model, as opposed to learning from examples [30]. Furthermore, we designed the prompt in a way that it can be applied across all types of tests and is robust to heterogeneity in the acceptance limits. It is structured to ask a simple "True" or "False" zero-shot question that is framed as follows: *"Evaluate the following electrical measure observation statement. Answer with just one "True" or "False" statement at the beginning of the answer. Is [measured_value] [acceptance_limits] ?"*. The LLM response is parsed, and the first "True" or "False" encountered is taken as the response, as sometimes the LLM might continue discussing and explaining the reasoning behind its decision.

Ontology-Based Data Access. We utilize a VKG to facilitate data access and manage structural heterogeneity. This VKG maps the validated Test Data storage to the ontology (refer to Figure 5). The knowledge within the virtualized semantic layer can be accessed via SPARQL queries, which are automatically translated into SQL. Listing 1.2 shows an RDF fragment modeling a POL Voltage Observation, which represents a row in the test table (refer to Figure 2). Each row is a `sosa:Observation` with a `sosa:hasSimpleResult` value. For instance, `<http://tasi.com/pol#TASI-1234-Core1>` is a `sosa:Observation` with a `sosa:hasSimpleResult` of "1.097 V". This observation is associated with the `sosa:observedProperty <POLVoltage>`. The SSN ontology has been extended with two properties to accommodate the specific needs of our case study. The `tasi:reportedIn` property links the observation to the corresponding Test Report, while the `tasi:hasAcceptanceLimits` property specifies the acceptable range for the observed property. For example, `tasi:hasAcceptanceLimits "[1.076, 1.224] V"` indicates that the acceptable voltage range for the POL Voltage Observation is between 1.076V and 1.224V. The `tasi:hasTestResult` property reports the "successful" value. For instance, a successful test is indicated by `tasi:hasTestResult "OK"`.

```
mappingId  POL_Voltage_Verification
target     tasi-pol:{tr_reference}-{v_cores} a sosa:Observation ;
           rdfs:label "{tr_reference}-{v_cores}";
           sosa:observedProperty tasi:POLVoltage;
           sosa:hasSimpleResult "{voltage_mesurements} V"^^cdt:ucum;
           tasi:hasAcceptanceLimits {acceptance_limits};
           tasi:testReportDate {test_report_date}^^xsd:dateTime;
           tasi:hasTestResult {successful};
           tasi:reportedIn {tr_reference}.
source     SELECT tr_reference, v_cores, voltage_mesurements
           acceptance_limits, test_report_date, successful
           FROM tasi.pol_voltage
```

Listing 3. The mapping for the `POLVoltage Observation`.

To populate the ontology, we establish a series of mappings. These mappings create connections between the ontology and the underlying data storage, thereby providing semantic meaning to the Test Data. An example of mapping for a `POLVoltageObservation` is provided in Listing 3. The mapping is defined with a `mappingId` of `POL Voltage Verification`, which corresponds to the type of test being performed. The `target` of the mapping is a URI that represents a `sosa:Observation` in the ontology. The `source` is a SQL query that retrieves the necessary data from the `POL Voltage Verification` table in the test results storage. The variables in the `source` query correspond to the placeholders in the `target`. Once the mapping is executed, these placeholders are replaced with the actual values retrieved by the `source` query. This allows us to virtually represent the storage as an RDF graph, integrating different data sources into a unified view.

6   Implementation and Evaluation

In this section, we delve into the specifics of our system's implementation and the technologies used. Following this, we present a benchmarking study of various state-of-the-art LLMs to evaluate their capability of performing automated compliance checking. An evaluation of the whole methodology from a cost-benefit perspective is provided in Section 7.

Implementation details. An *Apache Airflow* DAG (Directed Acyclic Graph) was designed to orchestrate the entire process. Apache Airflow is a popular open-source tool for creating, scheduling, and monitoring data pipelines. For modeling the Test Reports and their properties, we implemented the KG using *Apache Jena Fuseki*, a server that allows for querying and updating the KG using the SPARQL query language. The test results are stored in an *Apache Parquet*, a free and open-source column-oriented data storage, which allows handling large volumes of data while maintaining high performance. The VKG was implemented using *OntopSpark* [3], an extension developed by Politecnico di Milano of *Ontop* [31], an open-source OBDA system that allows for querying relational data sources through an ontology via R2RML [8] mappings. We do not report a detailed analysis of Ontop performances since it was benchmarked in several other papers [6, 15]. We present a discussion about the effort to solve the problem with and without our solution in Section 7.

LLMs Benchmarking. A benchmarking study was carried out to assess the performance of state-of-the-art LLMs in automated compliance checking. The models tested included GPT-3.5 [40], GPT-4 [25], Gemini Ultra [13], Mixtral 8x7B [14], LLama 2 70B [36], and Claude 3 Opus [2]. Performance was measured using standard metrics such as accuracy, precision, recall, and F1-score. The positive class was considered when the measured values fell outside the acceptance limit range, which is also the less represented class. The models were tested across three test categories: POL Voltage Verification, Internal Isolation, and External Isolation.

Table 1. The results of the comparative analysis on state-of-the-art LLMs for compliance checking, highlighting the superior performance of GPT-4 and Gemini Ultra.

Model	Test Type	#Tests	Accuracy	Precision	Recall	F1-Score
GPT-3.5	POL Voltage	53	0.868	0.625	0.556	0.588
	Internal Isolation	86	0.779	0.056	0.333	0.095
	External Isolation	59	0.932	0.333	0.333	0.333
	Overall	198	0.849	0.241	0.467	0.318
GPT-4	POL Voltage	53	0.981	1.000	0.900	0.947
	Internal Isolation	86	1.000	1.000	1.000	1.000
	External Isolation	59	0.983	0.750	1.000	0.857
	Overall	198	**0.990**	**0.938**	0.938	**0.938**
Gemini Ultra	POL Voltage	53	1.000	1.000	1.000	1.000
	Internal Isolation	86	1.000	1.000	1.000	1.000
	External Isolation	59	0.949	0.500	1.000	0.667
	Overall	198	0.985	0.833	**1.000**	0.909
Mixtral 8x7B	POL Voltage	53	0.925	0.875	0.700	0.778
	Internal Isolation	86	0.663	0.150	0.200	0.171
	External Isolation	59	0.644	0.136	0.600	0.222
	Overall	198	0.727	0.260	0.433	0.325
LLama 2 70B	POL Voltage	53	0.887	0.800	0.444	0.571
	Internal Isolation	86	0.733	0.091	0.400	0.148
	External Isolation	59	0.712	0.111	0.667	0.191
	Overall	198	0.768	0.178	0.471	0.258
Claude 3 Opus	POL Voltage	53	1.000	1.000	1.000	1.000
	Internal Isolation	86	0.895	0.250	1.000	0.400
	External Isolation	59	0.983	0.750	1.000	0.857
	Overall	198	0.949	0.615	**1.000**	0.762

Table 1 presents the benchmarking results, showing a clear distinction in performance among the tested LLMs. GPT-4 and Gemini Ultra are the top performers across all test categories, with GPT-4 achieving the highest overall accuracy, precision, and F1-score. Gemini Ultra, on the other hand, achieved perfect scores in the POL Voltage and Internal Isolation tests and had the highest overall recall. In contrast, GPT-3.5, Mixtral 8x7B, and LLama 2 70B consistently underperformed compared to the top models, rendering them unsuitable for our task. Claude 3 Opus demonstrated strong performance in the POL Voltage test but had lower precision in the Internal Isolation and External Isolation tests due to its inability to handle cases where the test result's unit of measure was absent, a frequent scenario in the Isolation tests. This benchmarking study provides strong evidence supporting the effectiveness of an LLM-based approach for automated compliance checking using state-of-the-art LLMs. The performance of GPT-4 and Gemini Ultra underscores the potential of these models in managing complex data validation tasks.

7   Discussion on Uptake and Lessons Learned

Benefits and Scalability. The transition from the current method to the proposed solution suggests a significant reduction in the effort measured in person-days required to complete and validate Test Reports before extracting longitudinal Test Data to analyze. The existing procedure necessitates substantial manual work for tasks such as creating Test Report templates, instantiating Test Reports, filling in the test results and checking the compliance with the acceptance limits, reviewing the Test Data and their coherence with the reported success, looking for errors and correcting them, and extract/transform data to perform longitudinal analysis (see Figure 3). The proposed solution, while requiring the modeling and maintenance of an ontology that encapsulates various test types (refer to Figure 1), and the annotation of the template with semantic tags that define each section, is fully automated (see Figure 4). This includes the extraction of test results and acceptance limits, error isolation, requests for manual review and correction, and data accessibility for longitudinal analysis.

Fig. 6. Comparison of effort reduction between our solution (KG+LLM) and the current method (AS-IS). The effort depends on the number of templates (n), the number of reports (x-axis), and the test types per report which for simplicity are set to 30.

We developed a cost model based on the experience documented in this paper. This model estimates the effort involved as a product of three factors: the number of different Test Report templates, the average number of Test Reports per template, and the number of test types (e.g., POL Voltage Verification, Preliminary Power Consumption) per report. Comparing the effort required to model, compile, and validate from 1 to 10 Test Report templates, each with an average of 30 types of test per report (refer to Figure 6), we derive that as the number of reports (on the x-axis) increases:

- For a single Test Report template (n=1), benefits start to appear after the 6th report.
- For five templates (n=5), benefits are seen after the 3rd report.
- For ten templates (n=10), the benefits are obtained at the 2nd report.

As the number of Test Report templates (n) increases, the number of Test Reports (on the x-axis) needed to see the benefits of using KG and LLMs is significantly reduced, with potential time savings of more than 50%. The right part of Figure 6 illustrates the break-even points for an increasing number of Test Reports in detail.

Next Steps for Large Scale Deployment. The proof-of-concept of the proposed solution was well received by Thales Alenia Space, but additional efforts are needed for the transition to a large-scale deployment. We are currently engaged in a feasibility study to port the solution to the other five nations in which Thales Alenia Space operates (France, Belgium, Spain, Switzerland, and the UK). Since the scenarios can vary significantly in these different divisions, this can potentially broaden adoption across the aerospace industry through further development and demonstration of value in diverse operational environments. Furthermore, despite the proposed solution focusing on a specialized case study, the principle of data validation via LLMs, to simplify the conceptual modeling process and reduce manual work, could potentially be extended to other scenarios such as mechanical and electrical qualification, given the ability of KGs and LLMs to adapt to different tasks and data types. However, it's true that to apply this approach in different contexts, slight reconfigurations would be necessary. Furthermore, it would be essential to establish benchmarks for each specific use case to evaluate the applicability in a new scenario. Given the significant savings, Thales Alenia Space expresses its intention to continue prototyping for other types of tests on PCBs, extend to the other product lines, and eventually deploy to all product lines. Preliminary experiments in this direction have already produced some promising results.

Lessons Learned. The development of the proposed solution revealed several key lessons. Firstly, the success of the implementation heavily relied on a well-structured ontology and clean mappings. The initial investment in modeling proved beneficial, as it minimized downstream efforts. Additionally, the integration of LLMs streamlined data validation, drastically reducing the need for manual intervention. Identified best practices include the necessity for iterative development and validation of the ontology and its corresponding mappings.

This ensures accurate modeling of test template reports. Moreover, it is crucial to conduct a comparative evaluation of alternative LLMs to stay updated with the evolving heterogeneities in Test Data, acceptance limits, and report requirements. Collaboration with stakeholders and domain experts was also essential for fine-tuning the KG and LLM prompts for optimal performance, and ensuring that confidentiality requirements were met while incorporating closed-source LLMs in the pipeline. As we move forward, these insights will guide our efforts to extend the solution to other product lines and further enhance the system's performance and reliability.

8   Conclusion and Future Work

In this paper, we demonstrated a successful application of Semantic Web technologies combined with LLMs for integrating and validating heterogeneous and unstructured industrial data through a use case related to Test Reports of electronic boards used in Thales Alenia Space's satellite systems. Our benchmarking study revealed that GPT-4 and Google Gemini possess remarkable abilities in automating the process of compliance checking. Considering that LLMs are still in the early stages of their development, it's reasonable to expect their performance to improve further, enabling them to handle even more complex data validation tasks in the near future. Overall, the proposed solution demonstrates a clear cost-benefit advantage over the existing document-centric solution. The potential efficiency gains underscore the value of investing in advanced AI-driven automation for such data-intensive tasks.

As future work, we intend to investigate whether the use of LLMs can be extended to perform automatic ontology construction, utilizing document tags, and also data homogenization. This would involve parsing the test results and the acceptance limits through an LLM-based approach. Additionally, we aim to further enhance data access by employing LLMs to convert natural language into SPARQL queries, thereby enabling Thales Alenia Space engineers to access knowledge directly using natural language.

References

1. Allen, B.P., Stork, L., Groth, P.: Knowledge Engineering Using Large Language Models. Transactions on Graph Data and Knowledge **1**(1), 3:1–3:19 (2023). https://doi.org/10.4230/TGDK.1.1.3, https://drops-dev.dagstuhl.de/entities/document/10.4230/TGDK.1.1.3
2. Anthropic: The claude 3 model family: Opus, sonnet, haiku (2024), https://paperswithcode.com/paper/the-claude-3-model-family-opus-sonnet-haiku
3. Belcao, M., Falzone, E., Bionda, E., Valle, E.D.: Chimera: a bridge between big data analytics and semantic technologies. In: The Semantic Web–ISWC 2021: 20th International Semantic Web Conference, ISWC 2021, Virtual Event, October 24–28, 2021, Proceedings 20. pp. 463–479. Springer (2021)
4. Brandt, S., Neuenstadt, C., Özçep, Ö., Pinkel, C., Zheleznyakov, D., Horrocks, I., Möller, R., Kharlamov, E., Jiménez-Ruiz, E., Kotidis, Y., Lamparter, S., Mailis, T., Svingos, C., Ioannidis, Y.: Ontology-based integration of streaming and static relational data with optique. In: Proceedings of the ACM SIGMOD International Conference on Management of Data. pp. 2109–2112 (2016). https://doi.org/10.1145/2882903.2899385
5. Charron, B., Hirate, Y., Purcell, D., Rezk, M.: Extracting semantic information for e-commerce. In: The Semantic Web – ISWC 2016: 15th International Semantic Web Conference, Kobe, Japan, October 17–21, 2016, Proceedings, Part II. p. 273–290. Springer-Verlag, Berlin, Heidelberg (2016). https://doi.org/10.1007/978-3-319-46547-0_27, https://doi.org/10.1007/978-3-319-46547-0_27
6. Chaves-Fraga, D., Priyatna, F., Cimmino, A., Toledo, J., Ruckhaus, E., Corcho, O.: Gtfs-madrid-bench: A benchmark for virtual knowledge graph access in the transport domain. Journal of Web Semantics **65**, 100596 (2020). https://doi.org/https://doi.org/10.1016/j.websem.2020.100596, https://www.sciencedirect.com/science/article/pii/S1570826820300354
7. Compton, M., et al.: The ssn ontology of the w3c semantic sensor network incubator group. Journal of Web Semantics **17**, 25–32 (2012). https://doi.org/https://doi.org/10.1016/j.websem.2012.05.003, https://www.sciencedirect.com/science/article/pii/S1570826812000571
8. Das, S., Sundara, S., Cyganiak, R.: R2RML: RDB to RDF mapping language. https://www.w3.org/TR/r2rml/ (2012)
9. FAIRsharing.org: Quantities, units, dimensions and types (qudt). https://doi.org/10.25504/FAIRsharing.d3pqw7 (2014)
10. Fernandez, R.C., Elmore, A.J., Franklin, M.J., Krishnan, S., Tan, C.: How large language models will disrupt data management. Proc. VLDB Endow. **16**(11), 3302–3309 (jul 2023). https://doi.org/10.14778/3611479.3611527, https://doi.org/10.14778/3611479.3611527
11. Ghidini, T.: Materials for space exploration and settlement. Nature Materials **17**(10), 846–850 (Sep 2018). https://doi.org/10.1038/s41563-018-0184-4
12. Golebiowska, J., Dieng-Kuntz, R., Corby, O., Mousseau, D.: Building and exploiting ontologies for an automobile project memory. In: Proceedings of the 1st International Conference on Knowledge Capture. p. 52–59. K-CAP '01, Association for Computing Machinery, New York, NY, USA (2001). https://doi.org/10.1145/500737.500749, https://doi.org/10.1145/500737.500749
13. Google: Gemini: A family of highly capable multimodal models (2023), https://arxiv.org/abs/2312.11805

14. Jiang, A.Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D.S., de las Casas, D., Hanna, E.B., Bressand, F., et al.: Mixtral of experts (2024), https://arxiv.org/abs/2401.04088

15. Kalaycı, E.G., Grangel González, I., Lösch, F., Xiao, G., ul Mehdi, A., Kharlamov, E., Calvanese, D.: Semantic integration of bosch manufacturing data using virtual knowledge graphs. In: Pan, J.Z., Tamma, V., d'Amato, C., Janowicz, K., Fu, B., Polleres, A., Seneviratne, O., Kagal, L. (eds.) The Semantic Web – ISWC 2020. pp. 464–481. Springer International Publishing, Cham (2020)

16. Kharlamov, E., Hovland, D., Jiménez-Ruiz, E., Lanti, D., Lie, H., Pinkel, C., Rezk, M., Skjæveland, M.G., Thorstensen, E., Xiao, G., Zheleznyakov, D., Horrocks, I.: Ontology based access to exploration data at statoil. In: Arenas, M., Corcho, O., Simperl, E., Strohmaier, M., d'Aquin, M., Srinivas, K., Groth, P., Dumontier, M., Heflin, J., Thirunarayan, K., Staab, S. (eds.) The Semantic Web - ISWC 2015. pp. 93–112. Springer International Publishing, Cham (2015)

17. Kharlamov, E., Mailis, T., Mehdi, G., Neuenstadt, C., Özgür Özçep, Roshchin, M., Solomakhina, N., Soylu, A., Svingos, C., Brandt, S., Giese, M., Ioannidis, Y., Lamparter, S., Möller, R., Kotidis, Y., Waaler, A.: Semantic access to streaming and static data at siemens. Journal of Web Semantics **44**, 54–74 (2017). https://doi.org/https://doi.org/10.1016/j.websem.2017.02.001, https://www.sciencedirect.com/science/article/pii/S1570826817300124, industry and In-use Applications of Semantic Technologies

18. Kharlamov, E., Mehdi, G., Savković, O., Xiao, G., Kalaycı, E.G., Roshchin, M.: Semantically-enhanced rule-based diagnostics for industrial internet of things: The sdrl language and case study for siemens trains and turbines. Journal of Web Semantics **56**, 11–29 (2019). https://doi.org/https://doi.org/10.1016/j.websem.2018.10.004, https://www.sciencedirect.com/science/article/pii/S1570826818300520

19. Kim, M., Wang, S.T., Ostrowski, D., Rychtyckyj, N., Macneille, P.: Technology outlook : Federated ontologies and industrial applications. International Journal of Semantic Computing **10**, 101–120 (03 2016). https://doi.org/10.1142/S1793351X1650001X

20. Laban, P., Kryściński, W., Agarwal, D., Fabbri, A.R., Xiong, C., Joty, S., Wu, C.S.: Llms as factual reasoners: Insights from existing benchmarks and beyond (2023)

21. Laranjeiro, N., Soydemir, S.N., Bernardino, J.: A survey on data quality: Classifying poor data. In: 2015 IEEE 21st Pacific Rim International Symposium on Dependable Computing (PRDC). pp. 179–188 (2015). https://doi.org/10.1109/PRDC.2015.41

22. Lefrançois, M., Zimmermann, A.: The unified code for units of measure in rdf: cdt:ucum and other ucum datatypes. In: Gangemi, A., Gentile, A.L., Nuzzolese, A.G., Rudolph, S., Maleshkova, M., Paulheim, H., Pan, J.Z., Alam, M. (eds.) The Semantic Web: ESWC 2018 Satellite Events. pp. 196–201. Springer International Publishing, Cham (2018)

23. Nasseri, M., Brandtner, P., Zimmermann, R., Falatouri, T., Darbanian, F., Obinwanne, T.: Applications of large language models (llms) in business analytics – exemplary use cases in data preparation tasks. In: Degen, H., Ntoa, S., Moallem, A. (eds.) HCI International 2023 – Late Breaking Papers. pp. 182–198. Springer Nature Switzerland, Cham (2023)

24. Norman, A., Das, S., Rohr, T., Ghidini, T.: Advanced manufacturing for space applications. CEAS Space Journal **15**(1), 1–6 (Jan 2023). https://doi.org/10.1007/s12567-022-00477-6

25. OpenAI: Gpt-4 technical report (2024), https://arxiv.org/abs/2303.08774
26. Ostrowski, D., Rychtyckyj, N., Macneille, P., Kim, M.: Integration of big data using semantic web technologies. pp. 382–385 (02 2016). https://doi.org/10.1109/ICSC.2016.101
27. Pan, S., Luo, L., Wang, Y., Chen, C., Wang, J., Wu, X.: Unifying large language models and knowledge graphs: A roadmap. IEEE Transactions on Knowledge and Data Engineering p. 1–20 (2024). https://doi.org/10.1109/tkde.2024.3352100, http://dx.doi.org/10.1109/TKDE.2024.3352100
28. Petersen, N., Halilaj, L., Grangel-González, I., Lohmann, S., Lange, C., Auer, S.: Realizing an rdf-based information model for a manufacturing company – a case study. In: d'Amato, C., Fernandez, M., Tamma, V., Lecue, F., Cudré-Mauroux, P., Sequeda, J., Lange, C., Heflin, J. (eds.) The Semantic Web – ISWC 2017. pp. 350–366. Springer International Publishing, Cham (2017)
29. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., Dean, J., Ghemawat, S.: Language models are unsupervised multitask learners. In: OSDI'04: Sixth Symposium on Operating System Design and Implementation. pp. 137–150 (2018)
30. Reynolds, L., McDonell, K.: Prompt programming for large language models: Beyond the few-shot paradigm. In: Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems. CHI EA '21, Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3411763.3451760, https://doi.org/10.1145/3411763.3451760
31. Rodríguez-Muro, M., Kontchakov, R., Zakharyaschev, M.: Ontology-based data access: Ontop of databases. In: Alani, H., Kagal, L., Fokoue, A., Groth, P., Biemann, C., Parreira, J.X., Aroyo, L., Noy, N., Welty, C., Janowicz, K. (eds.) The Semantic Web – ISWC 2013. pp. 558–573. Springer Berlin Heidelberg, Berlin, Heidelberg (2013)
32. Rojas Melendez, Julian Andres and Aguado, Marina and Vasilopoulou, Polymnia and Velitchkov, Ivo and Van Assche, Dylan and Colpaert, Pieter and Verborgh, Ruben: Leveraging semantic technologies for digital interoperability in the European railway domain. In: The Semantic Web – ISWC 2021. vol. 12922, pp. 648–664 (2021), http://doi.org/10.1007/978-3-030-88361-4_38
33. Rychtyckyj, N., Raman, V., Sankaranarayanan, B., Kuma, P.S., Khemani, D.: Ontology re-engineering: A case study from the automotive industry. AI Magazine **38**(1), 49–60 (Mar 2017). https://doi.org/10.1609/aimag.v38i1.2712, https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/2712
34. Tan, Z., Zhou, B., Zheng, Z., Savkovic, O., Huang, Z., Gonzalez, I.G., Soylu, A., Kharlamov, E.: Literal-aware knowledge graph embedding for welding quality monitoring: A bosch case. In: Payne, T.R., Presutti, V., Qi, G., Poveda-Villalón, M., Stoilos, G., Hollink, L., Kaoudi, Z., Cheng, G., Li, J. (eds.) The Semantic Web – ISWC 2023. pp. 453–471. Springer Nature Switzerland, Cham (2023)
35. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., Lample, G.: Llama: Open and efficient foundation language models (2023)
36. Touvron, H., et al.: Llama 2: Open foundation and fine-tuned chat models (2023)
37. Trajanoska, M., Stojanov, R., Trajanov, D.: Enhancing knowledge graph construction using large language models (2023)
38. Xiao, G., Calvanese, D., Kontchakov, R., Lembo, D., Poggi, A., Rosati, R., Zakharyaschev, M.: Ontology-based data access: A survey. In: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intel-

ligence, IJCAI-18. pp. 5511–5519. International Joint Conferences on Artificial Intelligence Organization (7 2018). https://doi.org/10.24963/ijcai.2018/777, https://doi.org/10.24963/ijcai.2018/777

39. Xiao, G., Ding, L., Cogrel, B., Calvanese, D.: Virtual Knowledge Graphs: An Overview of Systems and Use Cases. Data Intelligence **1**(3), 201–223 (06 2019). https://doi.org/10.1162/dint_a_00011, https://doi.org/10.1162/dint_a_00011

40. Ye, J., Chen, X., Xu, N., Zu, C., Shao, Z., Liu, S., Cui, Y., Zhou, Z., Gong, C., Shen, Y., Zhou, J., Chen, S., Gui, T., Zhang, Q., Huang, X.: A comprehensive capability analysis of gpt-3 and gpt-3.5 series models (2023)

41. Zhou, D., Zhou, B., Zheng, Z., Soylu, A., Cheng, G., Jimenez-Ruiz, E., Kostylev, E.V., Kharlamov, E.: Ontology reshaping for knowledge graph construction: Applied on bosch welding case. In: Sattler, U., Hogan, A., Keet, M., Presutti, V., Almeida, J.P.A., Takeda, H., Monnin, P., Pirrò, G., d'Amato, C. (eds.) The Semantic Web – ISWC 2022. pp. 770–790. Springer International Publishing, Cham (2022)

42. Zhou, X., Sun, Z., Li, G.: Db-gpt: Large language model meets database. Data Science and Engineering pp. 1–10 (01 2024). https://doi.org/10.1007/s41019-023-00235-6

43. Zhou, X., Zhao, X., Li, G.: Llm-enhanced data management (2024)