Using Large Language Models for Healthcare Data Interoperability: A Data Mediation Pipeline to Integrate Heterogeneous Patient-Generated Health Data and FHIR

Research Paper

Torben Ukena[1], Robin Wagler[1], and Rainer Alt[1]

[1]Leipzig University, Information Systems Institute, Leipzig, Germany
{torben.ukena,robin.wagler,rainer.alt}@uni-leipzig.de

Abstract. Integrating heterogeneous patient-generated health data effectively is pivotal for ensuring patient-centered care. This paper explores the potential of large language models (LLMs) to streamline this integration by reducing the labor-intensive ontology creation process. We propose a data mediator pipeline that combines an LLM with an output validation mechanism to transform diverse data formats into FHIR. Two prompt engineering strategies were evaluated for structuring wearable-derived sleep data for clinical use. Our results demonstrate that LLMs can generate valid FHIR representations, improving healthcare data interoperability. However, challenges remain in handling complex data structures requiring aggregation, affecting semantic accuracy. Future advancements should focus on refining LLMs' ability to process structured health data reliably, ensuring seamless clinical integration. Despite these challenges, LLMs present a promising approach to standardizing health data, ultimately enhancing patient-centered care and decision-making.

Keywords: FHIR, semantic interoperability, large language models, hospital information system, patient-generated health data

1 Introduction

The digital transformation of industries like healthcare (Wessel et al., 2021) promotes data-driven concepts like patient-centered care (Ologeanu-Taddei et al., 2023; Weissenfels et al., 2025). A key challenge in this transformation is ensuring interoperability across heterogeneous Hospital Information Systems (HIS) (Torab-Miandoab et al., 2023; Rachuba et al., 2024) and facilitating the seamless incorporation of Patient-Generated Health Data (PGHD) e.g. from wearable devices into clinical workflows (Sanders et al., 2016). Despite the growing consensus on the benefits of PGHD in enhancing patient care, its integration into HIS remains constrained by interoperability barriers (Khatiwada et al., 2024). Poor interoperability leads to fragmented health information, negatively impacting clinicians and patients. Incomplete or inconsistent data availability can result in misinformed medical decisions, hinder care coordination, and create obstacles for patient self-management (Dinh-Le et al., 2019). Additionally, data sustainability is becoming a key factor in healthcare, as PGHD should remain accessible throughout a

20th International Conference on Wirtschaftsinformatik
September 2025, Münster, Germany

patient's lifetime and, ideally, beyond (Jarvenpaa and Essén, 2023). In sum, interoperability challenges impose a significant financial burden, accounting for up to 25% of healthcare costs in the US and EU due to inefficient data exchange (Pidun et al., 2021).

To tackle the interoperability issue, it is important to distinguish between the four commonly recognized levels of interoperability (Ukena and Alt, 2024): (1) Technical interoperability refers to the basic ability of systems to exchange data through compatible technical infrastructures (Lilleng and Centre, 2005). (2) Syntactic interoperability ensures that data exchanged between systems follows a shared structure or format (Lilleng and Centre, 2005). (3) Semantic interoperability ensures that the meaning of exchanged data is preserved and interpreted consistently across systems, typically through shared ontologies or medical terminologies (Lilleng and Centre, 2005). (4) Organizational interoperability involves the alignment of institutional policies, legal agreements, and collaborative practices that enable effective data sharing between organizations (Adebesin et al., 2013).

To illustrate the interoperability challenge further, consider a patient with sleep problems who consults a physician: analyzing the PGHD could help the physician gain deeper insights into the patient's problems. For this purpose, the PGHD must be exchanged between the patient's wearables and the HIS in the clinical institution. A CSV export of the PGHD would ensure technical interoperability, representing the initial stage among the four levels of interoperability (Lilleng and Centre, 2005). Technical interoperability is usually less of an issue (Sunyaev et al., 2023), but even when data can be exchanged technically between the wearable and the HIS, that data remains unstructured and needs further processing before the physician can analyze it. If wearables could output the data in a structured format like JavaScript Object Notation (JSON) with predefined vocabulary and grammar, e.g. for the sleep start time, the second level of interoperability would be reached (Lilleng and Centre, 2005). However, the predefined fields may differ among different wearables, therefore the interpretation remains challenging. Converting the wearable data into established data standards in the healthcare sector like Fast Healthcare Interoperability Resources (FHIR) and using Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) codes to define the resources (Mildenberger et al., 2002) marks the third level of interoperability, as it ensures that information has the same meaning regardless of which system processes the data (Lilleng and Centre, 2005). The fourth level of interoperability is called organizational (Whitman et al., 2006) or pragmatic interoperability, as explained above. It describes the willingness and commitment of all involved organizations to collaborate (Adebesin et al., 2013). In the case of sleep data exchange, it would mean that wearable and HIS providers agree to cooperate. While levels 1-3 of interoperability mainly depend on technical aspects, the fourth level also depends on management aspects. To unlock the full potential of PGHD for patient-centered care, at least the third level of interoperability is required.

A common approach to ensure interoperability is standardization. The primary drawback of standardization lies in its practical implementation. In particular, the commitment to fully adopting comprehensive standards to ensure semantic interoperability (Scheer and Habermann, 2000). Despite a commitment to a comprehensive standard, the diversity of individual systems often leads to scenarios where the degree of specialization in the standard falls short of achieving semantic interoperability, as the standard permits too

much variation (Sunyaev et al., 2023). In other words, the agreed standards lack quality (Folmer et al., 2011).

2 Research Gap

Intelligent systems have the potential to automatically transform information from syntactical to semantic interoperability standards (Sunyaev et al., 2023). Using such interoperability support tools might be easier than having multiple vendors agree on a highly specialized standard. The transformation process between input and output within these tools may be seen as a mapping (Khan et al., 2014), based either on simple manual mappings or more advanced ontologies (Roussey et al., 2011; Grethe et al., 2009; Zaremba et al., 2008; Kawu et al., 2023). However, ontologies are domain-specific and have no generalization abilities. Their creation still involves much manual effort and requires domain knowledge (Jaulent et al., 2018). Due to the fast-paced wearable market, including all proprietary wearable standards into these ontologies remains complicated.

This leads to the question whether recent advancements in Artifical Intelligence (AI) technologies have the potential to contribute to semantic interoperability with regards to the integration of PGHD. This mapping task is comparable to translation in Natural Language Processing (NLP): In NLP, a German, English, and Spanish sentence represents the same meaning in a language-specific syntax. To achieve semantic interoperability, a translation application should transfer the meaning of a sentence into various languages by adapting the syntax. The task for the data mediator is similar: the system should transfer the meaning of information presented in a device-specific syntax into a standard for semantic interoperability. Transformer architectures have shown superior performance in NLP translation tasks (Vaswani et al., 2017). Large Language Models (LLMs) rely on the transformer architecture to build models to create textual data (OpenAI et al., 2024). Thus, this paper hypothesizes that LLMs can help to automate the mapping of PGHD into FHIR and formulates the following research question:

> **RQ**: To what extent can LLMs improve the integration of heterogeneous PGHD for patient-centered care?

To answer this research question, we combine an LLM with prompt engineering to build a pipeline that transforms PGHD from multiple wearables with heterogeneous data output into standardized FHIR, a highly specialized standard widespread in the healthcare industry, which supports a large information model and, therefore, aims at semantic interoperability (Leroux et al., 2017). To address hallucinations, the pipeline includes a validation step that ensures the LLMs output adheres to the formal requirements of the FHIR standard. Since the proposed pipeline utilizes only pre-trained components, no additional training data, such as fine-tuning, is required, allowing for immediate application with minimal effort.

3 Related Work

3.1 FHIR

FHIR is becoming a significant interoperability standard in healthcare (Ayaz et al., 2021). It aims to exchange healthcare information in a standardized and modular way (Williams et al., 2023). By using an HTTP-based Representational State Transfer (REST)-ful protocol and supporting different data representations such as XML or JSON, FHIR implements established standards at the technical and syntactic interoperability layers. To ensure semantic interoperability, it leverages terminology providers such as SNOMED CT or Logical Observation Identifiers Names and Codes (LOINC) and supports the representation of clinical concepts (Leroux et al., 2017). A main building block in FHIR is the observation resource, which is used to represent measurements or assertions about a subject. This can include vital signs, lab results, or PGHD such as sleep metrics. Each observation consists of key elements that enable the consistent representation and interpretation of observational data across systems, making it a valuable element for integrating PGHD into clinical workflows (Vorisek et al., 2022).

3.2 Interoperability in Healthcare

In practice, implementing an HIS for hospital A may be incompatible with the implementation for hospital B, even if both HISs are from the same vendor. Information exchange between heterogenous HIS is challenging, as information compatibility can not be granted (Sunyaev et al., 2023). Several approaches exist to translate information in a syntactic interoperability standard into semantically interoperability standards like FHIR. Using traditional Electronic Data Interchange (EDI) converter solutions has become a more advanced way to perform such translations. For example, Allocca et al. (2022) proposed a system that aims to translate guidelines on physical activity into an FHIR-compatible framework. Similarly, Pfaff et al. (2019) used ontologies and rules to transform clinical data into FHIR. Hong et al. (2019) introduced a pipeline to translate unstructured Electronic Heath Records (EHR) data into FHIR using mapping rules, normalization rules and an NLP-specific FHIR extension. To address the challenge of manual effort in creating ontologies and rules, Kamala et al. (2020) employed the Word2Vec architecture (Mikolov et al., 2013) to generate word embeddings from textual data. These word embeddings were then leveraged to analyze the similarity of words and find words with similar meanings. However, using Word2Vec has the drawback that the word embeddings do not represent contextual information about the usage of the word (Corrêa and Amancio, 2019), which may harm the potential of their approach.

3.3 LLMs for Interoperability

Applying LLMs for applications that convert textual input data into a desired semantic interoperability standard has already seen some contributions in research. For example, Yoon et al. (2024) used unstructured EHR records and translated them into FHIR. The proposed approach relied on role-prompting and achieved better results than a rule-based

translation. At the same time, using a LLM greatly decreases the reliance on manual labor compared to approaches based on handcrafted rules. Li et al. (2023a,b) expanded on this concept by employing more advanced prompt engineering techniques. They provided the LLM with task instructions, an FHIR template, four to five examples for the FHIR translation, a list with terminology codes and the input text which should be translated. For evaluation purposes, they validated the output of their LLM using the official FHIR validation checker. Their results attribute LLMs abilities to solve such standard translation tasks in general, which underlines our hypothesis that LLMs may be used to automate the mapping between sleep data and FHIR. Large-sized models achieved better results in their experimental setting than smaller models. Further, they found that even when terminology codes are provided, the LLM sometimes tends to hallucinate and use terminology codes which are invalid or do not exist. However, this work is focused only on EHR data. It does not consider wearable data, possibly including a larger variety as it underlies fewer regulations than medical products.

4 Architecture

4.1 Pipeline

Our pipeline can be found in Figure 1. For the LLM, we decided to use OpenAI's 4o mini model, the lightweight version of the 4o model, which offers a significantly lower cost per inference compared to the full-size 4o model ($0.150 compared to $5.00 for 1M input tokens). We hypothesize that introducing the FHIR-mediator and following the self-refinement strategy (see section 4.2) could bridge the performance gap between small and full-size models by keeping the costs low. However, our pipeline works in principle with all LLMs. Thus, it could also be implemented with a local-hosted LLM to have complete control over the data processing.

To address hallucinations, a general problem to LLMs (Tonmoy et al., 2024) and also explicitly mentioned in Li et al. (2023b) for the FHIR translation task, we integrate a FHIR-validator, which should trigger the LLMs self-refinement abilities (Du et al., 2024). The work of Madaan et al. (2023) inspired the idea of integrating a validation mechanism. It should ensure that the model output matches the FHIR standard and can be seen as a form of self-refinement through feedback and reasoning (Tonmoy et al., 2024). This validator checks if the format follows the FHIR-standard and also verifies that the terminology codes from providers like SNOMED CT or LOINC exist and that their displayed names are correct. This is important, as prior work found that LLMs sometimes tends to invent codes which do not exist (Li et al., 2023b). Given an input, the LLM creates the FHIR representation of the input and the FHIR-validator returns whether the input is valid. In the latter case, an error message is returned, specifying which part of the input did not match the FHIR standard and why. The LLM can then use this error message to correct its output. The validation mechanism returns only valid FHIR. Unlike agentic AI solutions, where LLMs have access to various tools and can decide which ones are needed to solve a given task, our FHIR validator is an integral, fixed component of the pipeline and is therefore applied to every request.

To create our application, we used the LangChain framework (Chase, 2022). The LLM was accessed using the OpenAI API. The validation mechanism itself is a function which uses an API to access the official FHIR-validator from HL7. This pipeline could be integrated as middleware or proxy, receiving PGHD in any format and outputting it in standardized FHIR, allowing seamless integration into the HIS.

Our data mediator pipeline starts with system input from our dataset (see section 5.1). The input may contain special characters, such as newlines or tabs, confusing the model. To address this, the input undergoes parsing, during which any unwanted tokens are removed. LangChain offers tools to do that directly. For clarity, the aforementioned parsing steps have been excluded from the pipeline visualization. Once the input is parsed correctly, it is used to construct the model input, respectively, the prompt. We used different prompting strategies explained in section 4.2. Based on the prompt, the LLM produces an FHIR output. In some cases, the model produces output that includes extraneous strings, resulting in an invalid JSON representation. Therefore, we apply the parsing mechanism again before the output is fed into the FHIR-validator. If the FHIR-validator returns that the output is valid FHIR, our pipeline returns the LLM's output, and the process is finished. If the LLM has not produced a valid FHIR, the corresponding error message is then used to construct a correction prompt. Based on this correction prompt, the LLM produces a new output, again fed into the FHIR-validator. We allow the model to use up to five iterations of this cycle. The pipeline cancels the translation process if the LLM cannot output a valid FHIR after five iterations. It stores the output with the information that no valid FHIR could be generated.

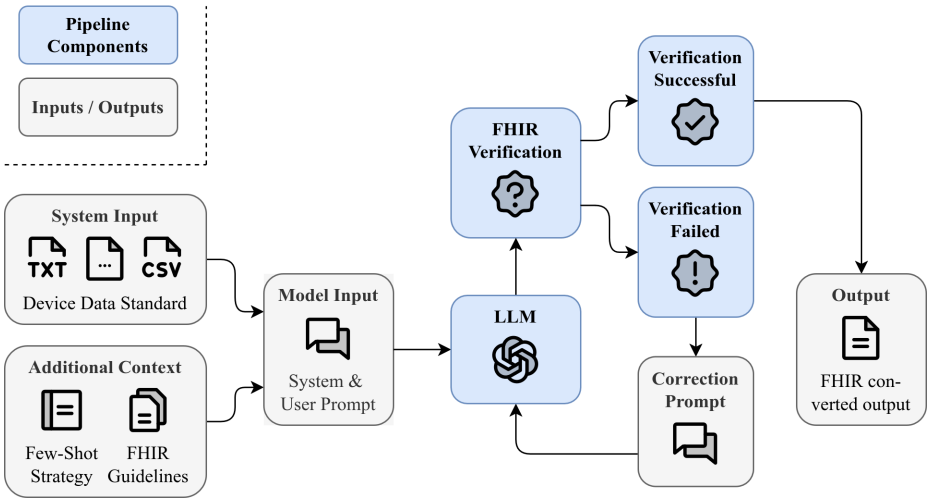

Figure 1. Architecture of proposed LLM-based data mediator.

4.2 Prompting Strategies

To tailor the pre-trained LLM to our task, we enhance the model's performance through prompt engineering. We deliberately avoid techniques that modify the initial weights of

LLMs, such as fine-tuning (Hu et al., 2022; Dettmers et al., 2023; Zhang et al., 2023) and minimize reliance on specialized training data, as obtaining such data is challenging in practice, particularly in critical domains like healthcare. This paper proposes and compares two prompt engineering strategies to solve the FHIR-translation task. The first strategy can be seen as a form of few-shot learning, where a few examples of input-output pairs are given to the model, so that it can quickly adapt to a new task (Parnami and Lee, 2022). As shown in Figure 2, the system prompt includes some general instructions about the task. In the user prompt, the model is asked to convert an input included in this prompt to FHIR. To match the few-shot paradigm, we provide the model with five examples of transforming different inputs into FHIR. The second strategy is based on reasoning (Huang and Chang, 2023) and inspired by the findings from (Yue Wu et al., 2023). Instead of providing the model with concrete examples of how a transformation to FHIR looks like, the model receives a guideline on transforming information into the FHIR standard. The system prompt includes these implementation guidelines and general role instructions. The user prompt remains the same as in the first strategy, except no transformation examples are provided. If the FHIR-validator drops an error, indicating that the model output did not meet the requirements for valid FHIR, we use a correction prompt to refine the model output. It mainly comprises general instructions and the corresponding terminology codes. Additionally, the prompt refers to the FHIR-validators error message.

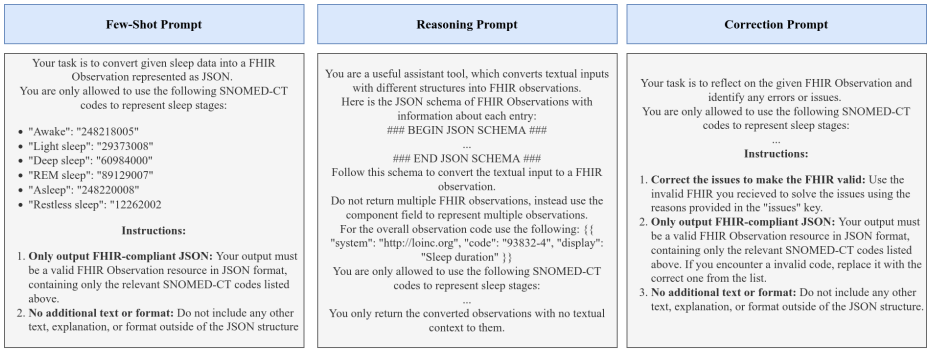

Figure 2. Prompting approaches integrated into the data mediator.

5 Data and Experiments

5.1 Data

To develop the concept and prototype for a LLM-based mediator that integrates between PGHD and HIS, this research has emerged from an applied project where sleep data is collected to improve sleep quality. During this project, subjects tracked their sleep and daylight activities using a Fitbit, Withings smart mattress, questionnaires, and multiple sensors over ten months. The sleep data in this context refers to measurements related to a user's sleep for a given period (usually a night). Typical examples for sleep data are

the different sleep phases like Wake, Light, Deep and REM sleep, as well as how long the user spends time in each phase or how often each phase occurs during the period. Even when most healthcare wearable market vendors offer the option to export their data as JSON or CSV, this is not sufficient for semantically interoperability. Consider the time a user spends awake during the night, called "wake" in the Fitbit syntax, while Withings designates the exact measurement as "wakeupduration". This issue illustrates that directly processing this data, e.g. into EHR, is nontrivial.

To the best of our knowledge, no public benchmark dataset is available for that kind of task. Therefore, we created our own dataset by randomly sampling 600 observations from the eSleepA[1] research projects' database. eSleepA aims to integrate multiple heterogenous data sources into a single system to provide sleep assistance. The database where we sampled our observations includes data collected form 20 participants, who submitted the data of different devices and questionnaires for 300 days each. The 600 samples in our dataset split as follows: 100 sleep observations from each Fitbit and Withings were represented in JSON format. The JSON formatted samples have the most complex structure among our datasets, as they either require combining multiple sleep observations of the same stage into one or mediating a lot of observations into FHIR. Additionally, we used 100 observations each to create CSV-formatted data. Further, we used 100 observations to create free-text TXT data and another 100 observations for Extensible Markup Language (XML) data, representing sleep observations in natural language, a known input structure for LLMs. The corresponding FHIR data was generated using a Python code with a pre-defined mapping. We validated the FHIR data using the official FHIR validator and reviewed it manually to ensure that this ground truth data is valid FHIR. Our dataset contains 600 sleep observations in different formats and their corresponding FHIR representation.

5.2 Experiments and Metrics

We propose the following task to evaluate the model's performance: Given sleep data as input, the model should translate them into FHIR. The task contains two challenges for the LLM: outputting valid FHIR and mapping the values from the input to the correct terminology codes and values. Sometimes, it may also be necessary to aggregate values from the input to a FHIR resource. While the performance for the first challenge depends on both the LLM and the FHIR validator, the performance for the second challenge depends purely on the LLM's capabilities to identify and map the values from the input correctly. To evaluate the performance of the data mediator, we defined the following experiments: Experiment E1 observes the performance of our pipeline in combination with the described few-shot strategy. For E2, we replace the few-shot strategy with the reasoning strategy. To assess the task-related performance of our model, we used the following metrics:

- **Validity Rate:** measures the proportion of the model's output that is valid FHIR. We checked the output using the official FHIR-validator. An output is considered valid if it follows the FHIR scheme and the terminology codes are valid. The validity rate

[1] https://leuris.uni-leipzig.de/portal/details/forschungsprojekt/8488

is defined as follows:

$$\text{Validity Rate} = \frac{\text{valid output}}{\text{total inputs}} \tag{1}$$

Given this definition, the best possible validation rate is 1.0 when all outputs are valid FHIR. Vice versa, a validity rate equal to 0.0 means that none of the outputs is valid FHIR. Additionally, we computed the validity rate for the first output of the model and denoted it with a . Thus, it represents the performance of the LLM without the FHIR-validator and the correction prompt. Therefore, it allows conclusions about the contribution of these components.

– **Semantic Accuracy Rate (SAR):** measures the proportion of the model's output that is semantically accurate. The SAR is assessed by comparing the model's output against our data set's ground truth FHIR representations. To do so, we compare the observations generated by our model with the observations in the ground truth. This ensures that the values for the observations are mapped and aggregated correctly and that all input measurements are transferred to the output. We defined semantic accuracy as follows:

$$\text{SAR} = \frac{\text{semantically accurate observations}}{\text{total observations}} \tag{2}$$

– **Correction Count:** represents how often the FHIR validator was called until the LLM produced valid FHIR output. If the correction count is 0.0, the LLMs output was valid FHIR instantly. A correction count of three means the validator was called three times, so the model's fourth output was valid FHIR. As mentioned earlier, we allowed the LLM up to five outputs so that the maximum correction count can be 4.0.

Note that the validity rate is computed over all model outputs. One model output usually contains multiple observations within the component object, so the other two metrics are computed per FHIR-resource. We run each experiment five times as the LLM output is not deterministic. If the FHIR-validator returns an error, we allow the model four additional shots to refine its output. If the model can not output valid FHIR after five iterations, the generation is halted and moved on to the next input.

6 Results and Discussion

6.1 Results E1

The results for experiment E1 are in table 1, and the corresponding standard deviation is shown in brackets behind the metric. We used the few-shot promoting strategy for experiment E1. When looking at the overall validity rate averaged over five runs, it can be found that our pipeline achieves a validity rate of 0.994 with a standard deviation of 0.079. As the validity rate is close to 1.0 (best possible result) and the standard deviation is relatively low (0.0786), we conclude that the few-shot strategy can output valid FHIR constantly. Observing the different input formats, we saw that the LLM struggles with translating Withings data in JSON format into FHIR. Using this data, the model achieved a validity rate of 0.790 and a relatively high standard deviation of 0.407. In terms of the

correction count, zero usage of the FHIR validator for both types of CSV inputs and the XML free-text inputs can be observed, indicating that the few-shot strategy produces valid FHIR within the first output for that input format. The correction count for Fitbit JSON and free-text TXT data shows that the model only calls the FHIR-validator in a few cases. When comparing the validity rate after the first output with the validity rate after the whole cycle, it is evident that using the FHIR validator significantly improves the validity rate and the standard deviation, as the model could correct its output based on the error message and the correction prompt. For Withings JSON, the correction count was 1.053 (standard deviation 1.646), which underlines our finding that the few-shot strategy struggles with that input format. Looking at the semantic accuracy draws a different picture: for both JSON inputs, our pipeline fails to produce semantically correct outputs. An explanation for that behaviour can be found in the structure of this input data: multiple measurements must be aggregated to create a semantic correct production. For example, the input data contains various values for REM sleep corresponding to the user's REM sleep phases during the night. This must be summed up to get the total REM sleep. The LLM fails to sum them up correctly, resulting in a semantically incorrect measurement. These results indicate that the JSON structure is too complex for our LLM as it fails to aggregate the values correctly. Similar results are observed for CSV data, although the model does not fail for all inputs, especially not for Fitbit data. The values were already aggregated in our textual inputs, resulting in a nearly optimal performance: All XML inputs were translated semantically correctly (semantic accuracy of 1.000), and only a few TXT inputs were processed semantically wrong (semantic accuracy of 0.989 with a standard deviation of 0.101). This underlines our verdict that the low performance on JSON and CSV inputs can be attributed to wrong aggregation due to the complex structure of the input data.

Table 1. Results E1

	Fitbit JSON	Withings JSON	Fitbit CSV	Withings CSV	free-text XML	free-text TXT	Overall
Validity Rate	1.000 (0.000)	0.790 (0.407)	1.000 (0.000)	1.000(0.000)	1.000 (0.000)	1.000 (0.000)	0.994 (0.079)
SAR	0.000 (0.000)	0.000 (0.000)	0.4943 (0.392)	0.024 (0.088)	1.000 (0.000)	0.989 (0.101)	0.631 (0.456)
Correction Count	0.005 (0.073)	1.053 (1.646)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.116 (0.301)	0.050 (0.358)
Validity Rate *	0.995 (0.073)	0.684 (0.467)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	0.884 (0.321)	0.972 (0.165)

6.2 Results E2

Table 2 presents the results using the reasoning strategy. As for E1, the corresponding standard deviation is shown in brackets. The reasoning strategy achieved a value of 0.818 under a standard deviation of 0.386 for the overall validity rate averaged over five runs. This is significantly lower than the results for the few-shot strategy. Further, the relatively high standard deviation indicates that the model struggles to perform constantly with the reasoning strategy. When looking deeper into the different input formats, it can be observed that the reasoning strategy works best with the free-text TXT inputs, as all outputs were valid FHIR, which aligns with the results for the few-shot strategy. Using free-text inputs in XML leads to a validity rate of 0.899, indicating that

the reasoning strategy can properly handle free-text inputs. We explain this behaviour because LLMs are trained primarily on free-text data in natural language and not on data with a proprietary structure like our sleep observations. Thus, they struggle to process such inputs correctly with complex proprietary structures. For the JSON inputs from both Withings and Fitbit, the averaged validity rate is 0.864, respectively 0.910. Translating CSV inputs into FHIR produces the worst results. A validity rate of 0.596 was achieved for the Fitbit CSV inputs, while the Withings CSV inputs led to a validity rate of 0.715. Therefore, we conclude that the reasoning strategy struggles with CSV inputs compared to the other input formats. Regarding the correction count overall inputs, the value of 1.552 indicates that the first output of the LLM is usually not valid FHIR; a standard deviation of 1.224 underlines this finding. When comparing the validity rate after the first output (0.059) with the validity rate after the correction (0.818), it is evident that our correction mechanism can significantly improve the quality of the output. This shows that even small size LLMs, which have lower operating costs, could refine their output based on the error message. In terms of semantic accuracy, the results with the reasoning strategy follow the direction of the results of the few-shot strategy. Our pipeline fails to aggregate the measurement from JSON completely. Nearly the same is true for the CSV inputs, even when some Fitbit measurements were aggregated correctly. Regarding the textual inputs, this strategy achieved a semantic accuracy of at least 0.585 for the XML input, significantly less than the few-shot strategy (1.000). For the TXT inputs, the reasoning strategy (semantic accuracy 0.984) is on par with the few-shot strategy (semantic accuracy 0.989). Overall, we conclude that the reasoning strategy lacks performance compared to the few-shot strategy. This underlines the hypothesis that LLMs do not have human-like reasoning skills. Instead, they rely on pattern matching (Schaeffer et al., 2023; Altmeyer et al., 2024; Mirzadeh et al., 2024; Wu et al., 2024; Webson and Pavlick, 2022; Lu et al., 2024).

Table 2. Results E2

	Fitbit JSON	Withings JSON	Fitbit CSV	Withings CSV	free-text XML	free-text TXT	Overall
Validity Rate	0.910 (0.286)	0.864 (0.344)	0.596 (0.491)	0.715 (0.452)	0.899 (0.302)	1.000 (0.000)	0.8180 (0.386)
SAR	0.000 (0.000)	0.000 (0.000)	0.015 (0.120)	0.000 (0.000)	0.585 (0.465)	0.984 (0.125)	0.138 (0.338)
Correction Count	1.347 (0.920)	1.173 (1.217)	2.331 (1.460)	1.848 (1.409)	1.292 (0.903)	1.047 (0.375)	1.552 (1.224)
Validity Rate *	0.008 (0.090)	0.255 (0.175)	0.032 (0.175)	0.053 (0.225)	0.008 (0.091)	0.000 (0.000)	0.059 (0.235)

7 Conclusion and Outlook

The experiments showed that even lightweight LLMs could have significant potential to translate PGHD in various formats into FHIR, therefore ensuring seamless interoperability between wearables heterogeneous HIS at low costs. The flexibility of LLMs marks a significant advantage over ontologies, as LLMs could adapt PGHD in various forms without extensive human effort. In terms of syntactical interoperability (second interoperability level), our data mediator showed a dedicated ability. Both strategies processed most of the inputs into valid FHIR; generally, the few-shot strategy performed better than the reasoning strategy. This indicates that clear transformation instructions

combined with examples are more beneficial for the LLM than a guideline of how FHIR should be created. Further, we showed that the validation mechanism, in combination with a correction prompt, contributes effectively to the performance of the data mediator, therefore underlining the self-refinement abilities of LLMs (Du et al., 2024). Regarding semantic accuracy (third level of interoperability), our data mediator failed to extract and aggregate measurements from complex input structures in nested JSON or large CSV, as the inputs were not processed correctly. However, using free-text data, the results showing a significantly higher semantic accuracy. This is likely because free-text data is more straightforward to process, as it does not require aggregation. Adding a preprocessing step that aggregates the measurements before they are fed into the model, could help to mitigate this issue. However, this would increase the manual effort to apply the pipeline, as aggregation rules for all possible inputs would be required. Another solution could be the usage of an LLMs with enhanced reasoning abilities, so that the LLMs understands which measurements belong together. However, these complexities seem to be beyond the reasoning capabilities of existing LLMs (Schaeffer et al., 2023; Altmeyer et al., 2024; Mirzadeh et al., 2024; Wu et al., 2024; Webson and Pavlick, 2022; Lu et al., 2024).

To return to our initial research question of to what extent LLMs can improve the integration of heterogeneous health data, our findings indicate that LLMs hold significant potential in enabling seamless health data standardization. The results on free-text data demonstrated that LLMs can effectively translate diverse health data into FHIR, making it more accessible for clinical use and patient self-management. The free-text results further suggest that LLMs may possess generalized capabilities for ensuring semantic interoperability beyond the healthcare domain, as LLMs can process inputs based on their contextual understanding, enabling adaptive data integration. However, challenges remain in handling structured data formats that require complex aggregation, which may impact usability for healthcare providers. Enhancing the processing of proprietary data structures and refining measurement aggregation mechanisms is crucial to fully unlocking the benefits of LLM-driven health data integration but also for maintaining data sustainability. Remarkably, our pipeline does not require expert knowledge or extensive training data, so it could be applied immediately to various HIS, mitigating the need for highly specialized ontologies.

An important direction for future research is the effective handling of complex data structures, which may require preprocessing steps such as measure aggregation. We plan to investigate whether integrating an LLM with function-calling capabilities into our pipeline can enhance the processing of such data. In particular, the use of deterministic functions (e.g., a calculator tool or a specialized aggregation tool) could improve the LLM's aggregation capabilities without adding additional manual effort and thus support more accurate processing of complex data. Beyond the mentioned area, we also plan to investigate whether some form of uncertainty estimation can be integrated into our pipeline to increase the quality of the output.

References

Adebesin, F., Foster, R., Kotzé, P., and van Greunen, D. (2013). A review of interoperability standards in e-health and imperatives for their adoption in Africa. *South African Computer Journal*, 50.

Allocca, C., Jilali, S., Ail, R., Lee, J., Kim, B., Antonini, A., Motta, E., Schellong, J., Stieler, L., Haleem, M. S., Georga, E., Pecchia, L., Gaeta, E., and Fico, G. (2022). Toward a symbolic AI approach to the WHO/ACSM physical activity & sedentary behavior guidelines. *Applied Sciences*, 12(4):1776.

Altmeyer, P., Demetriou, A. M., Bartlett, A., and Liem, C. C. S. (2024). Position: Stop making unscientific AGI performance claims. In Salakhutdinov, R., Kolter, Z., Heller, K., Weller, A., Oliver, N., Scarlett, J., and Berkenkamp, F., editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 1222–1242. PMLR.

Ayaz, M., Pasha, M. F., Alzahrani, M. Y., Budiarto, R., and Stiawan, D. (2021). The fast health interoperability resources (FHIR) standard: Systematic literature review of implementations, applications, challenges and opportunities. *JMIR Medical Informatics*, 9(7):e21929.

Chase, H. (2022). LangChain.

Corrêa, E. A. and Amancio, D. R. (2019). Word sense induction using word embeddings and community detection in complex networks. *Physica A: Statistical Mechanics and its Applications*, 523:180–190.

Dettmers, T., Pagnoni, A., Holtzman, A., and Zettlemoyer, L. (2023). QLoRA: Efficient finetuning of quantized LLMs. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Dinh-Le, C., Chuang, R., Chokshi, S., and Mann, D. (2019). Wearable health technology and electronic health record integration: Scoping review and future directions. *Jmir Mhealth and Uhealth*.

Du, Y., Wei, F., and Zhang, H. (2024). Anytool: Self-reflective, hierarchical agents for large-scale api calls.

Folmer, E., Luttighuis, P. O., and van Hillegersberg, J. (2011). Do semantic standards lack quality? A survey among 34 semantic standards.

Grethe, J. S., Ross, E., Little, D., Sanders, B., Gupta, A., and Astakhov, V. (2009). Mediator infrastructure for information integration and semantic data integration environment for biomedical research. *Methods Mol Biol*, 569:33–53.

Hong, N., Wen, A., Shen, F., Sohn, S., Wang, C., Liu, H., and Jiang, G. (2019). Developing a scalable FHIR-based clinical data normalization pipeline for standardizing and integrating unstructured and structured electronic health record data. *Jamia Open*, 2(4):570–579.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. (2022). Lora: Low-rank adaptation of large language models. In *ICLR*. OpenReview.net.

Huang, J. and Chang, K. C.-C. (2023). Towards reasoning in large language models: A survey.

Jarvenpaa, S. L. and Essén, A. (2023). Data sustainability: Data governance in data infrastructures across technological and human generations. *Information and Organization*, 33(1):100449.

Jaulent, M.-C., Leprovost, D., Charlet, J., and Choquet, R. (2018). Semantic interoperability challenges to process large amount of data perspectives in forensic and legal medicine.

Kamala, M., Hu, Y., Sigwele, T., Naveed, A., Sigwele, T., Kamala, M., and Susanto, M. (2020). Addressing semantic interoperability, privacy and security concerns in electronic health records. *Journal of Engineering and Scientific Research*, 2.

Kawu, A. A., Hederman, L., Doyle, J., and O'Sullivan, D. (2023). Patient generated health data and electronic health record integration, governance and socio-technical issues: A narrative review. *Informatics in Medicine Unlocked*, 37:101153.

Khan, W. A., Khattak, A. M., Hussain, M., Amin, M. B., Afzal, M., Nugent, C., and Lee, S. (2014). An adaptive semantic based mediation system for data interoperability among health information systems. *Journal of Medical Systems*, 38(8):28.

Khatiwada, P., Yang, B., Lin, J.-C., and Blobel, B. (2024). Patient-generated health data (pghd): Understanding, requirements, challenges, and existing techniques for data security and privacy. *Journal of Personalized Medicine*, 14(3).

Leroux, H., Metke-Jimenez, A., and Lawley, M. (2017). Towards achieving semantic interoperability of clinical study data with FHIR. *Journal of Biomedical Semantics*.

Li, Y., Wang, H., Yerebakan, H., Shinagawa, Y., and Luo, Y. (2023a). Enhancing health data interoperability with large language models: A fhir study.

Li, Y., Wang, H., Yerebakan, H. Z., Shinagawa, Y., and Luo, Y. (2023b). Fhir-gpt enhances health interoperability with large language models. *medRxiv*.

Lilleng, J. and Centre, B. (2005). Towards semantic interoperability. In *Proceedings of the Ifip/Acm Sigapp Interop-esa Conference*.

Lu, S., Bigoulaeva, I., Sachdeva, R., Madabushi, H. T., and Gurevych, I. (2024). Are emergent abilities in large language models just in-context learning?

Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegreffe, S., Alon, U., Dziri, N., Prabhumoye, S., Yang, Y., Gupta, S., Majumder, B. P., Hermann, K., Welleck, S., Yazdanbakhsh, A., and Clark, P. (2023). Self-refine: Iterative refinement with self-feedback.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space.

Mildenberger, P., Eichelberg, M., and Martin, E. (2002). Introduction to the DICOM standard. *European Radiology*, 12(4):920–927.

Mirzadeh, I., Alizadeh, K., Shahrokhi, H., Tuzel, O., Bengio, S., and Farajtabar, M. (2024). Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models.

Ologeanu-Taddei, R., Guthrie, C., and Jensen, T. B. (2023). Digital transformation of professional healthcare practices: fitness seeking across a rugged value landscape. *European Journal of Information Systems*, 32(3):354–371.

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., Bello, I., Berdine, J., Bernadett-Shapiro, G., Berner, C., Bogdonoff, L., Boiko, O., Boyd, M., Brakman,

A.-L., Brockman, G., Brooks, T., Brundage, M., Button, K., Cai, T., Campbell, R., Cann, A., Carey, B., Carlson, C., Carmichael, R., Chan, B., Chang, C., Chantzis, F., Chen, D., Chen, S., Chen, R., Chen, J., Chen, M., Chess, B., Cho, C., Chu, C., Chung, H. W., Cummings, D., Currier, J., Dai, Y., Decareaux, C., Degry, T., Deutsch, N., Deville, D., Dhar, A., Dohan, D., Dowling, S., Dunning, S., Ecoffet, A., Eleti, A., Eloundou, T., Farhi, D., Fedus, L., Felix, N., Fishman, S. P., Forte, J., Fulford, I., Gao, L., Georges, E., Gibson, C., Goel, V., Gogineni, T., Goh, G., Gontijo-Lopes, R., Gordon, J., Grafstein, M., Gray, S., Greene, R., Gross, J., Gu, S. S., Guo, Y., Hallacy, C., Han, J., Harris, J., He, Y., Heaton, M., Heidecke, J., Hesse, C., Hickey, A., Hickey, W., Hoeschele, P., Houghton, B., Hsu, K., Hu, S., Hu, X., Huizinga, J., Jain, S., Jain, S., Jang, J., Jiang, A., Jiang, R., Jin, H., Jin, D., Jomoto, S., Jonn, B., Jun, H., Kaftan, T., Łukasz Kaiser, Kamali, A., Kanitscheider, I., Keskar, N. S., Khan, T., Kilpatrick, L., Kim, J. W., Kim, C., Kim, Y., Kirchner, J. H., Kiros, J., Knight, M., Kokotajlo, D., Łukasz Kondraciuk, Kondrich, A., Konstantinidis, A., Kosic, K., Krueger, G., Kuo, V., Lampe, M., Lan, I., Lee, T., Leike, J., Leung, J., Levy, D., Li, C. M., Lim, R., Lin, M., Lin, S., Litwin, M., Lopez, T., Lowe, R., Lue, P., Makanju, A., Malfacini, K., Manning, S., Markov, T., Markovski, Y., Martin, B., Mayer, K., Mayne, A., McGrew, B., McKinney, S. M., McLeavey, C., McMillan, P., McNeil, J., Medina, D., Mehta, A., Menick, J., Metz, L., Mishchenko, A., Mishkin, P., Monaco, V., Morikawa, E., Mossing, D., Mu, T., Murati, M., Murk, O., Mély, D., Nair, A., Nakano, R., Nayak, R., Neelakantan, A., Ngo, R., Noh, H., Ouyang, L., O'Keefe, C., Pachocki, J., Paino, A., Palermo, J., Pantuliano, A., Parascandolo, G., Parish, J., Parparita, E., Passos, A., Pavlov, M., Peng, A., Perelman, A., de Avila Belbute Peres, F., Petrov, M., de Oliveira Pinto, H. P., Michael, Pokorny, Pokrass, M., Pong, V. H., Powell, T., Power, A., Power, B., Proehl, E., Puri, R., Radford, A., Rae, J., Ramesh, A., Raymond, C., Real, F., Rimbach, K., Ross, C., Rotsted, B., Roussez, H., Ryder, N., Saltarelli, M., Sanders, T., Santurkar, S., Sastry, G., Schmidt, H., Schnurr, D., Schulman, J., Selsam, D., Sheppard, K., Sherbakov, T., Shieh, J., Shoker, S., Shyam, P., Sidor, S., Sigler, E., Simens, M., Sitkin, J., Slama, K., Sohl, I., Sokolowsky, B., Song, Y., Staudacher, N., Such, F. P., Summers, N., Sutskever, I., Tang, J., Tezak, N., Thompson, M. B., Tillet, P., Tootoonchian, A., Tseng, E., Tuggle, P., Turley, N., Tworek, J., Uribe, J. F. C., Vallone, A., Vijayvergiya, A., Voss, C., Wainwright, C., Wang, J. J., Wang, A., Wang, B., Ward, J., Wei, J., Weinmann, C., Welihinda, A., Welinder, P., Weng, J., Weng, L., Wiethoff, M., Willner, D., Winter, C., Wolrich, S., Wong, H., Workman, L., Wu, S., Wu, J., Wu, M., Xiao, K., Xu, T., Yoo, S., Yu, K., Yuan, Q., Zaremba, W., Zellers, R., Zhang, C., Zhang, M., Zhao, S., Zheng, T., Zhuang, J., Zhuk, W., and Zoph, B. (2024). Gpt-4 technical report.

Parnami, A. and Lee, M. (2022). Learning from few examples: A summary of approaches to few-shot learning.

Pfaff, E. R., Champion, J., Bradford, R. L., Clark, M., Xu, H., Fecho, K., Krishnamurthy, A., Cox, S., Chute, C. G., Overby Taylor, C., and Ahalt, S. (2019). Fast healthcare interoperability resources (FHIR) as a meta model to integrate common data models: Development of a tool and quantitative validation study. *JMIR Medical Informatics*, 7(4):e15199.

Pidun, U., Knust, N., Kawohl, J., Avramakis, E., and Klar, A. (2021). The untapped potential of ecosystems in health care. *Boston Consulting Group*.

Rachuba, S., Reuter-Oppermann, M., and Thielen, C. (2024). Integrated planning in hospitals: a review. *OR Spectrum*.

Roussey, C., Pinet, F., Kang, M. A., and Corcho, O. (2011). Ontologies for interoperability. In *Ontologies in Urban Development Projects*, pages 39–53. Springer London, London.

Sanders, J. P., Loveday, A., Pearson, N., Edwardson, C. L., Yates, T., Biddle, S. J. H., and Esliger, D. W. (2016). Devices for self-monitoring sedentary time or physical activity: A scoping review. *Journal of Medical Internet Research*.

Schaeffer, R., Miranda, B., and Koyejo, S. (2023). Are emergent abilities of large language models a mirage?

Scheer, A.-W. and Habermann, F. (2000). Enterprise resource planning: Making ERP a success. *Communications of the ACM*, 43(4):57–61.

Sunyaev, A., Dehling, T., Strahringer, S., Da Xu, L., Heinig, M., Perscheid, M., Alt, R., and Rossi, M. (2023). The future of enterprise information systems. *Business & Information Systems Engineering*, 65(6):731–751.

Tonmoy, S. M. T. I., Zaman, S. M. M., Jain, V., Rani, A., Rawte, V., Chadha, A., and Das, A. (2024). A comprehensive survey of hallucination mitigation techniques in large language models.

Torab-Miandoab, A., Samad-Soltani, T., Jodati, A., and Rezaei-Hachesu, P. (2023). Interoperability of heterogeneous health information systems: A systematic literature review. *BMC Medical Informatics and Decision Making*.

Ukena, T. and Alt, R. (2024). Contributions of AI to advance interoperability with data mediators. In *Wirtschaftsinformatik 2024 Proceedings*, volume 117.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Vorisek, C. N., Lehne, M., Klopfenstein, S. A. I., Mayer, P. J., Bartschke, A., Haese, T., and Thun, S. (2022). Fast healthcare interoperability resources (fhir) for interoperability in health research: Systematic review. *JMIR Med Inform*, 10(7):e35724.

Webson, A. and Pavlick, E. (2022). Do prompt-based models really understand the meaning of their prompts?

Weissenfels, S., Nissen, A., and Smolnik, S. (2025). Advancing digital health in information systems research: Insights from a text mining analysis. *Electronic Markets*, 35(1):23.

Wessel, L., Baiyere, A., Ologeanu-Taddeï, R., Cha, J., and Jensen, T. B. (2021). Unpacking the difference between digital transformation and it-enabled organizational transformation. *J. Assoc. Inf. Syst.*, 22:6.

Whitman, L. E., Panetto, H., and Desilva, D. (2006). The missing link: Culture and language barriers to interoperability. *14746670*, 39(4):51–57.

Williams, E., Kienast, M., Medawar, E., Reinelt, J., Merola, A., Klopfenstein, S. A. I., Flint, A. R., Heeren, P., Poncette, A.-S., Balzer, F., Beimes, J., von Bünau, P., Chromik, J., Arnrich, B., Scherf, N., and Niehaus, S. (2023). A standardized clinical data harmonization pipeline for scalable AI application deployment (FHIR-DHP): Validation and usability study. *JMIR Medical Informatics*.

Wu, Z., Qiu, L., Ross, A., Akyürek, E., Chen, B., Wang, B., Kim, N., Andreas, J., and Kim, Y. (2024). Reasoning or reciting? exploring the capabilities and limitations of language models through counterfactual tasks.

Yoon, D., Han, C., Kim, D. W., Kim, S., Bae, S., an Ryu, J., and Choi, Y. (2024). Redefining health care data interoperability: Empirical exploration of large language models in information exchange. *Journal of Medical Internet Research*, 26:e56614.

Yue Wu, Shrimai Prabhumoye, So Yeon Min, Yonatan Bisk, Ruslan Salakhutdinov, Amos Azaria, Tom Mitchell, and Yuanzhi Li (2023). SPRING: Studying the paper and reasoning to play games.

Zaremba, M., Herold, M., Zaharia, R., and Vitvar, T. (2008). Data and process mediation support for b2b integration. volume 359.

Zhang, Q., Chen, M., Bukharin, A., He, P., Cheng, Y., Chen, W., and Zhao, T. (2023). Adaptive budget allocation for parameter-efficient fine-tuning. In *The Eleventh International Conference on Learning Representations*.