Data & Knowledge Engineering 152 (2024) 102313

Contents lists available at ScienceDirect

Data & Knowledge Engineering

journal homepage: www.elsevier.com/locate/datak

To prompt or not to prompt: Navigating the use of Large Language Models for integrating and modeling heterogeneous data

Adel Remadi [a], Karim El Hage [a,*], Yasmina Hobeika [a], Francesca Bugiotti [a,b]

[a] *CentraleSupélec, 3 Rue Joliot Curie, Gif-Sur-Yvette, 91190, Île De France, France*
[b] *Laboratoire Interdisciplinaire des Sciences du Numérique, Université Paris-Saclay, CNRS, Orsay, 91400, Île De France, France*

ARTICLE INFO

Keywords:
Data engineering
Large language models
Conceptual schema modeling
Entity resolution
Data integration
Property graph models

ABSTRACT

Manually integrating data of diverse formats and languages is vital to many artificial intelligence applications. However, the task itself remains challenging and time-consuming. This paper highlights the potential of Large Language Models (LLMs) to streamline data extraction and resolution processes. Our approach aims to address the ongoing challenge of integrating heterogeneous data sources, encouraging advancements in the field of data engineering. Applied on the specific use case of learning disorders in higher education, our research demonstrates LLMs' capability to effectively extract data from unstructured sources. It is then further highlighted that LLMs can enhance data integration by providing the ability to resolve entities originating from multiple data sources. Crucially, the paper underscores the necessity of preliminary data modeling decisions to ensure the success of such technological applications. By merging human expertise with LLM-driven automation, this study advocates for the further exploration of semi-autonomous data engineering pipelines.

1. Introduction

Data integration is a critical step of any pipeline when considering multiple heterogeneous data sources [1]. Our previous work [2] has shown how an interconnected graph schema, modeled on Neo4j using data sources of different structures and languages, can yield insights that would not have otherwise existed had the sources existed in the database independently. However, the manual construction of the graph database presented a significant limitation: the full integration was arduous and time-consuming. Despite considerable advancements in data integration automation, both through traditional semantic techniques [3,4] and recent language model applications [5], there remains a critical dependency on extensive fine-tuning over large training datasets. The necessity for extensive training stems from the requirement for models to possess a deep comprehension of linguistic subtleties and domain-specific knowledge relevant to the studied use-case [5].

Large Language Models (LLMs) have significantly enhanced the ease with which we can retrieve and interpret data, showcasing the ability to handle a diverse range of tasks. LLMs often require merely one or a few examples to perform tasks, and in certain cases, have outperformed traditional supervised models in terms of effectiveness and efficiency [6,7]. Despite the potential benefits, [8] points out that the effectiveness of LLMs in data integration, especially in completing complex tasks like entity matching or resolution, remains uncertain. On the other hand, [9] argues that the unique ability of LLMs to understand semantic ambiguities and integrate data from real-world scenarios necessitates a fundamental rethinking of established data management approaches. This perspective underscores the necessity to recognize the potential benefits of incorporating these advanced tools into data management

[*] Corresponding author.
E-mail addresses: adel.remadi@student-cs.fr (A. Remadi), karim.elhage@student-cs.fr (K. El Hage), yasmina.hobeika@student-cs.fr (Y. Hobeika), francesca.bugiotti@centralesupelec.fr (F. Bugiotti).

https://doi.org/10.1016/j.datak.2024.102313
Received 28 February 2024; Received in revised form 1 May 2024; Accepted 1 May 2024
Available online 6 May 2024

strategies. Considering this, our paper investigates the use of LLMs to aid in the automation of data extraction and integration tasks. The work further investigates the collaborative role that human data modeling design could play to enhance such automated pipeline. This is done by designing a conceptual schema for a unique and heterogeneous dataset from scratch, elaborating on the importance of the design considerations. Consequently, we were able to use the schema to both guide the prompts fed into the LLM and ensure that the output of the LLM respects the proposed schema. As a result, this paper demonstrated that the use of LLMs, guided by prompts that consider human data modeling considerations, is a very encouraging approach to automate the integration of data originating from heterogeneous sources. Hence, the contributions of this work are as follows:

1. Introducing a conceptual schema methodology designed to accommodate a selected dataset composed of multiple sources, each varying in format and language.
2. Automating the extraction of entities from unstructured data sources using a Large Language Model in the context of the defined conceptual schema.
3. Automating the data integration of entities originating from multiple data sources (structured/unstructured data) using a Large Language Model in the context of the defined conceptual schema.

The introduction of data modeling and integration using LLMs into this paper's methodology not only addresses the manual and time-consuming aspects of traditional data integration processes, but also addresses the advanced capabilities of LLMs to understand and process language nuances. This approach enables a more efficient and effective integration of diverse data sources, widening the range of possibilities for data integration practices in various fields.

The remainder of the paper is structured as follows. Section 2 introduces the related work that supports the different approaches and strategies considered in our scientific methodology. Section 3 introduces the different types of structured and unstructured sources that are used in our study. Section 4 details the data modeling choices that served as a foundation for the integration of the heterogeneous sources and the manner in which an LLM can be used to automate the data integration process. Section 5 assesses the quality of the proposed automated data integration process and describes some key takeaways and implications. Finally, Section 6 summarizes our findings and proposes possible avenues for future research.

2. Related work

Our approach lies in creating a graph representation of data coming from different sources to enable the execution of predictive Artificial Intelligence algorithms [2]. Achieving this objective requires appropriate data engineering considerations, including the definition of a conceptual model to help design, develop and run these artificial intelligence solutions [10–12]. New research fields are opening strong opportunities for the definition of conceptual models [13,14]. Simultaneously, research has also been devoted to addressing data integration with novel approaches [15–17]. As a result, it is imperative to investigate advancements in both fields: modeling and integration. Prior to that, since the implementation is performed in Neo4j, a portion of this section is dedicated to further understanding the benefits of using such a tool.

2.1. Data modeling using Neo4j graphs

One of the key challenges of our study has been to integrate various sources of information of different structures and languages. For example, [18] already considered that integrating diverse and complex information such as structured databases, unstructured text, and multimedia content represented a significant challenge in Big Data applications. NoSQL databases have been discussed as an appropriate solution for such endeavors due to their ability to adapt to different sources and data formats, as well as their high-performance capabilities and enhanced flexibility [19].

Graph data structures, which belong to the NoSQL family, are applied in areas where information about data inter-connectivity or topology is of importance [20]. Modeling data as graphs allows querying relationships in the same manner as querying the data itself. Instead of calculating and querying the connection steps as in relational databases, graph databases read the relationship from storage directly [20]. Neo4j employs the so-called Property Graph Model [21]. Like any other graph database model, it relies on two types of entities: nodes and edges. However, Property Graphs contrast with other graph data models in the way that they allow the storing of properties directly on nodes and edges [21], which is not the case for other graph data models such as RDF [22]. Recent literature [23] commented on how graph databases are easily scalable, fast, efficient, and flexible. This was confirmed by [24] that utilized Neo4j to model a time-evolving social network. The objective was to capture human activities and interactions sourced from mobile devices and wearable sensors. Notably, the study showcases the effectiveness and scalability of real-world queries, highlighting the efficiency of the approach [24]. Our study capitalizes on the capabilities of Neo4j to establish a directed graph, facilitating the visualization of pertinent insights. The choice of Neo4j was particularly interesting, as it offered us the abilities to take advantage of the interconnectedness of a graph structure, while handling different data sources in a flexible and integrated manner.

2.2. Conceptual modeling and artificial intelligence

Datasets are nowadays analyzed by algorithms and systems with growing complexity. Conceptual modeling has always been instrumental in understanding data and complex systems. For decades, the research community has dedicated large attention to modeling and dug in topics that include data modeling, process modeling, meta modeling, and model quality [25–27]. One of the main questions during the last few years has been: "how conceptual modeling can help structure machine learning and practitioners' projects ?" [11,28,29]. The conclusion has been that machine learning and data modeling can complement and help each other [30] even to the point of defining systems that can auto-configure and optimize themselves [31]. The attention around this topic is increasing to the point that a new research area identified with the CMAI acronym (Conceptual Model and Artificial Intelligence) recently started to be developed [32]. In a similar vein, our conceptual modeling work has been oriented to complement value adding AI applications, such as the recommender system proposed in our previous study [2]. Our present paper adopts a reciprocal approach by taking advantage of Large Language Models to enhance data engineering tasks.

2.3. Advances in data modeling and integration

Defining a good conceptual model is still an open challenge in many research areas. Even recent literature shows how a big research community is still working on defining and validating conceptual models for use-cases such as smart homes [33], European laws [34] or even manufacturing business analytics [35]. Similarly, studies have shown that defining a conceptual model that integrates many heterogeneous data sources is an even more complex and open challenge [27,36]. Many open questions persist, particularly in the context of new tools and approaches like LLMs [16,17] or the synergy between knowledge graphs and natural language [37].

The last several years have seen significant efforts to explore the use of NLP techniques and applications of language models in the context of databases systems and conceptual modeling [38–40]. These applications also include data discovery and integration [16,41,42]. For example, very encouraging results have emerged in using GPT-3.5 for the task of entity extraction from unstructured documents [43,44]. Other works such as [13,14] propose LLM-based tools that extract document values from data lakes. Recent research has also focused on considering GPT-3 in support of model construction and definition [45] or data transformation [46]. Some studies even attempted to substitute databases and data models with Generative AI Machines [47].

The problem of data integration has been widely studied in literature [15,48]. Classical solutions traditionally define a unified framework based on general meta-structures and a set of rules to map the sources into a target model [49,50]. In a similar fashion, our work maps all the available data into a target schema made of entities coming from different data sources. According to our research, a conceptual model is indeed essential to succeed in integrating data from heterogeneous sources. That is why, our present study explores how LLMs can be used to support the automation and enrichment of a graph data model. This research field is only starting to be explored, but some approaches have already shown good results [13,14,43,45].

3. Dataset

The dataset used to demonstrate the data modeling and integration methodology is the same one used in our previous work [2]. The research falls under the Vrailexia project, an EU-funded project comprised of a consortium of universities across Europe [51]. The three different data sources made available as part of the project were questionnaires, interview transcripts, and virtual reality (VR) simulations. The content of this data centers around learning disorders in higher education. Hence, the details in this section shall be heavily specific to this topic. Each source will be described to provide the context for the data modeling considerations in Section 4.

3.1. Questionnaire

The Vrailexia project has collected valuable data from dyslexic and non-dyslexic students through questionnaires digitally distributed in high schools and universities in France and Spain.

Data description

The questionnaires capture the perception of students with respect to how potential difficulties affect them in their studies and how useful they would consider specific tools/strategies to cope with these challenges. Hence, the data collected from this source are purely personal subjective opinions of the respondent. The questionnaires are provided in tabular form, serving as the first structured data source available for use. Table 1 describes the data source's structure and its main components. The questionnaire collects personal information relating to the respondents such as age, gender, dyslexic members in family, and educational background. Furthermore, it aims to understand what are the respondents' potential learning disorders, learning difficulties and their perceived usefulness of tools and learning strategies.

Table 1
Breakdown of questionnaire columns.

Category	Number of columns
Personal Information	45
Learning Disorders	6
Severity of Learning Difficulties (Scale 1–5)	13
Usefulness of Tools (Scale 1–5)	18
Usefulness of Learning Strategies (Scale 1–5)	22

Some of the tools and learning strategies are filled with the answer ''I don't know'' to indicate that a student was not familiar with a specific solution (see examples in Appendix Table A.4). There were a total of 2106 respondents collected from both France and Spain. Approximately 23% of the respondents needed to be discarded as a result of leaving the majority of fields blank. 16% of the respondents had Dyslexia, often combined with other learning disorders. It proved difficult to collect data for a large percentage of dyslexic respondents given that Dyslexia affects 5–17.5% of the population [52,53]. The average age of respondents is 21.5 and the majority are Female (69.5%). The average rated severity across all problems by students with learning disorders is 3.16 compared to 2.43 for students without any learning disorder.

Data pre-processing

The pre-processing of the French and Spanish questionnaires involved several steps to clean and transform the data. The transformed columns were renamed to be more concise and descriptive. These final names would eventually be used as the names of the nodes modeled in the graph database. For example, questions such as "What is your age? Do not enter your date of birth" and "Which university are you from?" are reformulated to "Age" and "University" respectively. Some columns such as the age required some additional pre-processing as the answer formats were not consistent or were invalid. Overall, these pre-processing steps helped to clean and organize the questionnaire data, ensuring that it was in a suitable format for graph creation in Neo4j.

3.2. Virtual reality (VR) simulations

As part of the present project, data collection from VR simulations was performed with dyslexic students and non-dyslexic students. The purpose of the VR test is to investigate whether providing Dyslexic students with an interactive and immersive setting could enhance their learning experience, whilst also educating teachers on the considerations to make for students in such condition [54].

Data description

Today, data has been collected in French, Spanish, and Italian universities. The participants are asked to perform two types of tests in a VR environment: (1) A Silent Reading test to assess performance; (2) A Psychometric Rosenberg [55] test for the assessment of anxiety, self-esteem and self-efficacy. The silent reading portion of the VR is a text comprehension exercise of which a respondent has to answer a series of elementary questions based on a text. The psychometric portion of the test (Rosenberg) seeks to survey the respondent's level of confidence by asking them to rank a series of general questions on a scale of 1 (Strongly Disagree) to 5 (Strongly Agree).

The data are exported in tabular format in three separate tables each storing the information about the user, silent reading test, and Rosenberg test respectively. For example, the table storing the data regarding the silent-reading contains two columns for each of the six questions: the first column, a boolean representing whether the respondent answered the question correctly, and the second, the time that has elapsed (in seconds) since the beginning of the test upon the respondents completing their answer. A sample of a few columns from the three tables have been joined in Appendix Table A.5.

Overall, the data from the VR provide a complementary secondary structured data source with information about the respondents that would need to be integrated with data from the Questionnaire. At the time of conducting this research, only 100 responses were collected using the VR technology (of which 40% had at least one learning disorder) as it was only rolled-out for data collection in 2023. Hence, the VR data was only used to model the schema of the database and to show how data from different sources can be integrated.

Data pre-processing

As the first crucial pre-processing step, the names of respondents were anonymized by dropping the information for analysis. In the silent-reading test, response times were recorded in a cumulative manner each time a respondent answered a question. As the property of interest was the elapsed time for each individual question, the cumulative time records were transformed accordingly. In this test, the respondents' disorders were all collected in one column and so the answers needed to be parsed such that each disorder was label-encoded. Finally, the age column required similar pre-processing as that described in the questionnaire by correcting answers provided in an invalid format.

3.3. Interview transcripts

The interviews' data was made available as text files of text-to-speech transcribed questions and answers with 10 French experts.

Data description

Various topics were covered related to students with learning disorders. Broadly speaking, the content of the interviews could be extracted and categorized into several themes: the learning disorders cited in each interview, problems encountered by students with such disorders, and finally the tools/strategies that could be useful in coping with learning disorders or a specific difficulty. There were roughly 25 questions per interview. The interview transcripts serve as the unstructured data source in demonstrating the methodology.

Data pre-processing

No pre-processing was conducted on these files. However, the title of the page was removed and names of the experts and interviewers were anonymized by replacing them with “Expert” and “Interviewer” respectively.

4. Methodology

This section describes the methodological steps undertaken to model and integrate data from the multiple sources of interest. In Section 4.1, the modeling approach is introduced and later demonstrated with a conceptual schema of the structured and unstructured data sources. After that, Section 4.2 details the automation of data extraction. Section 4.3 addresses how the extracted entities were disambiguated. Finally, Section 4.4 describes the data resolution of instances originating from the different data sources.

The methodological steps of our study, as detailed in Section 4, are structured to enable the adaptation of our modeling approach to a broad array of use-cases. As mentioned previously in Section 1, and further justified in 2.1, Neo4j is the database system of choice to store information from the various data sources. Hence, the conception of the schema is conducted in a manner that follows the conventions of graph data modeling and Neo4j design. This means that schema representations are in property graph model form whereas queries are demonstrated using Cypher: a query language optimized for property graphs [56].

4.1. Modeling the conceptual schema

Conceptual models offer the ability to integrate heterogeneous sources, creating a base for uncovering insights, and developing data-driven solutions. However, designing such conceptual models that deal with multiple sources can present multiple challenges. There are key differences in structure, format, and content between the sources as well as differences that may exist within each source itself. The following Sections 4.1.1 and 4.1.2 both describe our data modeling steps and introduce the pillars that compose the final schema of the integrated and interconnected graph database (Fig. 6).

4.1.1. Structured data sources

As described in more details in Section 3, the questionnaire collects the following information about the respondents:

- their personal information
- their learning disorders, if any
- their self-assessment about how problems associated with the disorders affect them in their daily lives
- their perception of the usefulness of tools and learning strategies that are known to be used by students with learning disorders.

Given the central role of the respondents, we decided to model them as nodes containing their personal information as properties (e.g. anonymized identifier, age, and gender). Learning disorders (e.g. dyslexia, dysorthographia, dyspraxia, etc.) could also be treated as characteristics of respondents but were instead modeled as an independent node type, since there was an interest in capturing their relations to other nodes. Each problem, tool, and strategy was then categorized under their own respective node types as they interact with the respondent rather than being inherent characteristics. Under these modeling choices, each respondent was linked to the other four defined node types. Hence, the corresponding schema was centered around a dedicated node type called `Respondent`, as shown in Fig. 1.

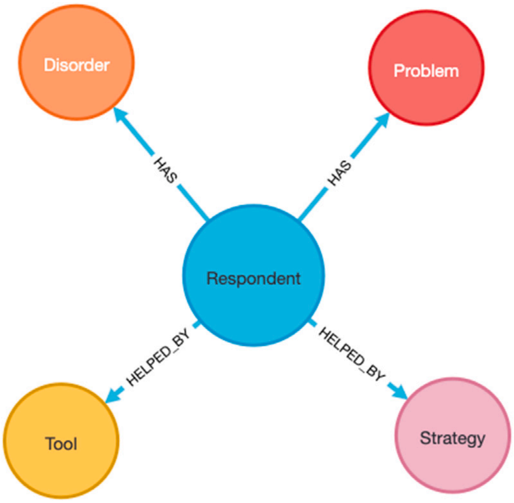


Fig. 1. Conceptual schema of the relationship between the `Respondent` node and the nodes derived from the questionnaire.

A `Respondent` node *HAS* a set of `Disorder` nodes and a set of `Problem` nodes. The `Respondent` node is also *HELPED_BY* sets of `Strategy` and `Tool` nodes. The relationships between the different nodes had to consider the answers of the respondent, who rated each problem, tool, and strategy on a scale from one to five — a measure of a respondent's connection with a specific node. These values were modeled as the Strength attribute of the relationship. As an example, to find the respondents who consider certain problems to be the most severe, one could use a navigation scheme making use of the Strength edge attribute. In Cypher syntax:

```
(: Respondent)-[: HAS {strength: 5}]->(:Problem)
```

The answers of each respondent are easily traced back thanks to this representation. It was decided to relate every respondent to all the nodes of types `Problem`, `Tool`, and `Strategy`, irrespective of the strength of their answer. The one exception was in the case where the answer was left blank as this meant that the respondent had no prior experience or knowledge about the concerned instance.

Storing the Strength attribute in the relationships instead of in the `Problem`, `Tool`, and `Strategy` (PST) nodes ensures that no information is lost and prevents node or attribute redundancies. This modeling choice further facilitates the use of graph science algorithms to process the database as a weighted graph. One limiting consequence however is that clustering algorithms, such as k-means, are restricted to node attributes in Neo4j [57]. In our schema, executing these functions would imply disaggregating edge properties (such as Strength), defeating the purpose of their modeled intent.

The VR data source, that serves as the second source of structured data, complements the questionnaire by providing further details about the characteristics of the respondents. The VR test is modeled in a similar way by creating relationships between `Respondent` nodes and the two additional node types (`Test` and `Confidence`), as illustrated in Fig. 2. Answers from the silent reading test were modeled under `Test` nodes, while responses to the Rosenberg test fell under `Confidence` nodes. The two consequent relationships depict the cases where a respondent *ANSWERED* a test that measured their reading performance and *FEELS* a specific confidence level, as indicated through the Rosenberg questions. Similar to the Questionnaire, the test results and confidence level of the respondent are stored as an edge attribute. For example, if the goal was to identify respondents who answered a question correctly in less than ten seconds, the Cypher query for extracting the information from the graph is:

```
(: Respondent)-[r: ANSWERED {correct-answer: True})->(:Test)
WHERE r.time-taken < 10
```

The provided conceptual schema allows for a natural integration between the Questionnaire and VR test through the `Respondent` and `Disorder` nodes. In practice, it is important to consider that there are issues that require additional attention before achieving true integration. One such issue is the multi-lingual nature of the dataset (the questionnaire existed in both French and Spanish). Node names were stored in French by default but also had complementary attributes with machine translations in English, Spanish, and Italian. There is a long history of studies on the effectiveness of commercialized Neural Machine Translation Models such as Google Translate to translate in many languages across several applications [58,59]. Another issue is to deal with cases where a respondent had contributed to both the questionnaire and VR tests. Some controls were implemented to address such

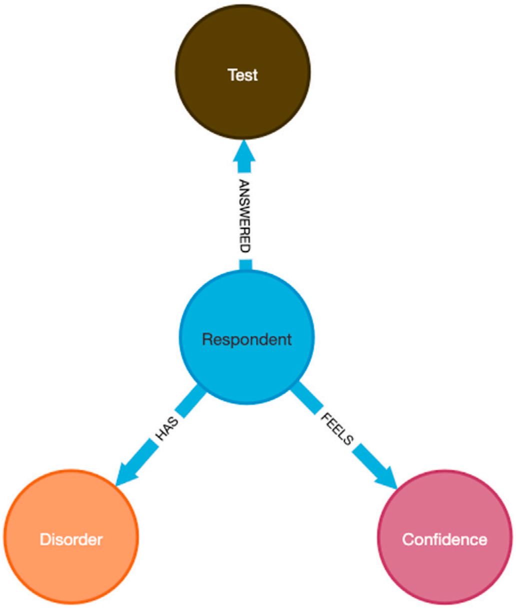


Fig. 2. Conceptual schema of the relationship between a `Respondent` node with nodes derived from VR dataset.

cases to prevent any overwriting of personal information, thus ensuring proper and accurate data reconciliation. An additional attribute named Source was created within `Respondent` nodes to trace which data sources a respondent completed. Integrating these structured data sources was eventually rather straightforward thanks to the properties that were introduced for reconciliation. In contrast, the task was considerably more challenging for the unstructured data coming from the interviews.

4.1.2. Unstructured data sources

Modeling unstructured data sources is significantly less intuitive than that of their structured counterparts. Whereas structured data nodes and relationships can be intuitively interpreted, unstructured data sources require more complex considerations. Moreover, relying solely on human assessment could hinder any attempt to automate the data engineering pipeline. As described in Section 3, the unstructured data of this study was collected in the form of interview transcripts with experts to better understand the characteristics of learning disorders, the problems they may cause in higher education, and the ways in which affected students could address these problems. Using this information, it is possible to enrich the existing schema by modeling a new node type, `Expert`, that is critical for tracing the source of stored interview data. Each expert is modeled as a node with a unique anonymized identifier, having the language of the interview and the name of the transcript file stored as attributes of the node. Fig. 3 highlights the schema modeled with this new node type.

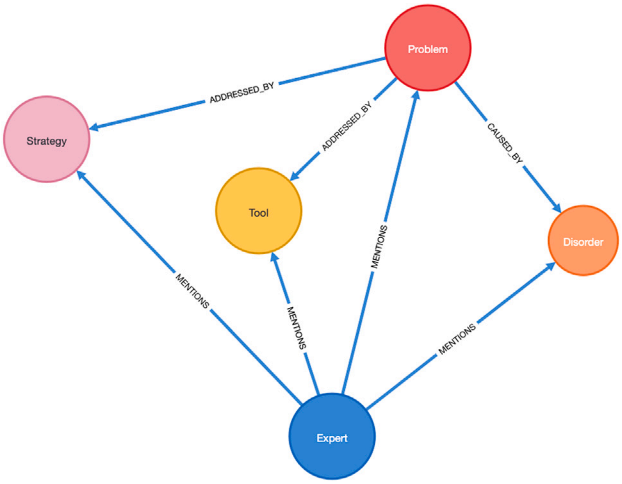


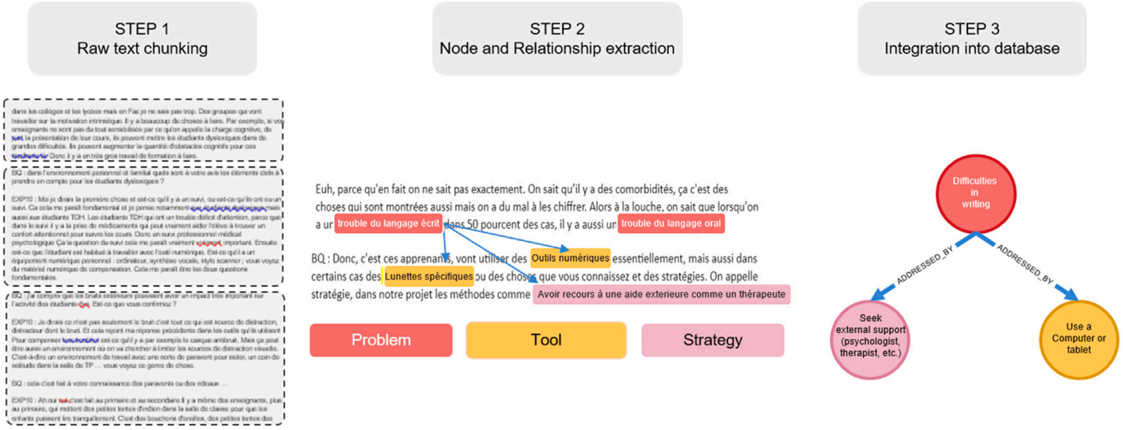
Fig. 3. Conceptual schema of the relationship between the `Expert` node, the nodes derived from the interview, and the resulting causal relationships.

The `Expert` node, as such, *MENTIONS* other nodes. This enables both to trace the origin of any `Disorder` or `PST` nodes to their source expert. There are two additional relationships that can be further inferred from the interview data, namely that `Problem` nodes can be *CAUSED_BY* specific `Disorder` nodes and that `Problem` nodes can be *ADDRESSED_BY* `Tool` and `Strategy` nodes. These relationships are critical in that they create causal links between the different nodes, hence contributing to a more interconnected graph structure. Moving forward, a Named-Entity Recognition task was designed and implemented to efficiently extract data from the interview transcripts in an automated and scalable fashion. Its aim was to automatically extract information from transcripts according to the modeling decisions illustrated in Fig. 3. The following subsection shall detail the methodology employed for this step of the data engineering pipeline.

4.2. Data extraction

As the structured sources are available in tabular form, categorizing the columns into their respective nodes is sufficient for loading data into the database. Specific transformations are made to facilitate the loading of such data into Neo4j, but these are not to be detailed as they are not the focus of this paper. In contrast, the data of interest from the unstructured sources are not immediately accessible. In the example of the interviews, the data relating to each entity is scattered throughout the transcripts. Therefore, a Named Entity Recognition (NER) task is required before any database integration. An illustration of the task to be performed is proposed in Fig. 4. To ensure the scalability of the data integration pipeline, it is imperative to rely on an automated method to conduct the extraction process. Our paper proposes to do so using Open AI's "GPT-3.5-Turbo" Large Language Model (LLM). As previously discussed, this approach aims to demonstrate that with the correct data modeling choices and prompting, there is a promising path to automating data integration in a generalized manner without necessarily requiring heavy machine learning model training and deployment. The NER task conducted by the LLM needs to be able to perform node and relationship extraction like the one illustrated in Fig. 4.

As part of the process of extracting nodes and relationships, unstructured interview transcripts underwent a series of processing steps. First, these sources were segmented into manageable chunks to accommodate the LLM's context window — its input token limit. Each chunk was composed of a sequence of a question from the interviewer followed by the corresponding expert's answer.



Euh, parce qu'en fait on ne sait pas exactement. On sait qu'il y a des comorbidités, ça c'est des choses qui sont montrées aussi mais on a du mal à les chiffrer. Alors à la louche, on sait que lorsqu'on a ur trouble du langage écrit dans 50 pourcent des cas, il y a aussi ur trouble du langage oral

BQ : Donc, c'est ces apprenants, vont utiliser des Outils numériques essentiellement, mais aussi dans certains cas des Lunettes spécifiques ou des choses que vous connaissez et des stratégies. On appelle stratégie, dans notre projet les méthodes comme Avoir recours à une aide exterieure comme un thérapeute

Fig. 4. NER-based integration of interview data.

Chunks were all assigned metadata containing their file name and exact location in the raw transcripts. As stated in Section 4.1.2, such information is later stored as node attributes to ensure the traceability of each piece of information.

Second, a prompt was constructed to employ the LLM to conduct the NER task. The drafted prompt provides details on the information to be extracted as well as formatting guidelines. Here, it was important to provide context in a manner that respected the schema previously shown in Fig. 3. The prompt also incorporates a few-shot learning approach by feeding the LLM with example chunks along with the respective nodes and relationships that can be extracted from them. Constructing a strong prompt was particularly challenging, as the LLM can be prone to hallucinate or deviate from its specified task. There is no established comprehensive method yet to evaluate prompt design [60]. Hence, our prompt engineering step required many iterations and refinements to cope with the sensitivity of the LLM's interpretation of its provided instructions. The significant role of prompt optimization to improve results was also demonstrated in other studies [61]. An excerpt from our final prompt can be found in Appendix Fig. B.7.

Third, a rigorous post-processing pipeline was implemented to ensure the proper formatting of the extracted entities. The LLM outputs were formatted strings of texts, on which a series of controls were applied to ensure their conformity to the prompted instructions. Outputs were transformed into lists of nodes and relationships that were consequently loaded into the graph database. These entities were only introduced into the database if they respected the modeled schema in Fig. 3. All imported nodes names were stored in their original language. Machine translations of these names were added as node attributes in all the other official languages of the Vrailexia project. This task was done as part of the NER process to ensure that the translations account for the context used by the LLM during extraction. New studies have already shown the competitiveness of LLMs at translation compared to traditional approaches [62,63]. Our NER method enabled dealing with unstructured data in an automated way, the quality of which is further addressed in Section 5. Prior to that, a complementary task to NER in charge of handling extracted duplicates, called disambiguation, is described in the next subsection.

4.3. Node disambiguation

A disambiguation strategy was deployed as the final processing step of the unstructured data extraction to enhance data representation and trim out "near" duplicate node names from the NER task described in Section 4.2. The disambiguation involves computing textual embeddings of the node names and their pairwise cosine similarity values. The node names were first pre-processed to remove stop words and frequent words specific to each node type prior to computing these embeddings. Nodes with a cosine similarity of 0.98 and higher are flagged for merging. The duplicate candidates are consequently merged together by selecting one node name to be kept. Ideally, the preserved node name is the one having the most number of pairs in the duplicate groups. In cases where multiple nodes held the highest number of duplicates, the preserved name was randomly selected from among them. As this step aims to identify duplicates, it was reasonable, through trial and error, to set such a high threshold of cosine similarity.

Node disambiguation was not a focus of this paper but rather a sub-step between NER and entity resolution. The decision to further explore or optimize disambiguation in the future shall be made depending on the outcomes of these two steps. Nevertheless, disambiguation helped to ensure that the database does not suffer from a large volume of redundancies, which is essential to the data integration described in the next subsection.

4.4. Data integration and resolution

The final step of the data engineering pipeline involves the integration of the dataset in a manner that enables navigation across multiple data sources. Specifically, data resolution (or entity matching/entity resolution) is achieved by connecting similar

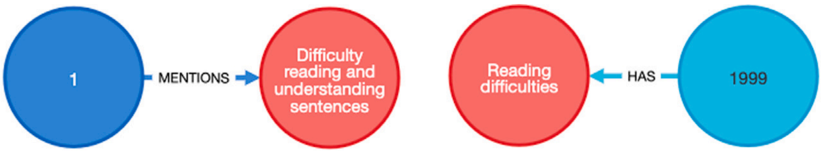


Fig. 5. Example of two syntactically similar Problem node names originating from the questionnaire and interview transcripts respectively.

Problem, Tool, and Strategy nodes coming from heterogeneous sources. The illustrative example in Fig. 5 depicts two Problem nodes coming from different sources conveying synonymous meanings. Data resolution aims to connect those two nodes. Such an operation was critical in our previous work [2], which aimed at developing a recommender system use-case based on a multi-source graph database. Data resolution was required to recover the insights from the expert interviews on how to address the most severe issues of Dyslexic students from the questionnaires. This task had however been previously handled manually in [2], requiring a considerable amount of time and representing an obstacle for automation.

Automating this approach faces a challenge: the inherent synonymy across the data sources is not always as explicit as illustrated in Fig. 5. Similar nodes are often connected through analogous descriptions, contexts, or situations. Simply considering the cosine similarity of textual embeddings or resorting to other traditional semantic approaches is insufficient to capture such nuanced similarities [64] without introducing many false positives and false negatives. Therefore, it was necessary to take on the difficult endeavor of not only resolving nodes that had syntactic similarities as that shown in Fig. 5, but also resolving nodes having a contextual or nuanced common meaning, such as between "Reading Difficulties" and "Size of Text" (a relationship thematic in nature). An LLM prompting approach was again employed to systematically attempt the challenge of achieving data integration in an autonomous manner.

Several attempts were made at engineering a prompt that provided the LLM with sufficient context to label a pair of nodes. The final prompt defined that nodes would be linked if they shared one of three types of similarities: syntactic, thematic, or functional. Each similarity type was carefully defined in the prompt. In addition, the model was asked to explain the reason behind deeming a pair similar or not before providing a label. Research has shown that such chain-of-thought prompting could improve the ability of LLMs to conduct complex tasks [65]. The excerpt from our final prompt can be found in Appendix Fig. B.8. An *IS_SIMILAR* relationship is introduced into the schema in cases where node names within the same node type are deemed similar and happen to originate from different data sources. The final conceptual schema therefore ensures interconnectedness, enabling comprehensive data analyses. Fig. 6 shows the final schema including the data resolution provided by this latest step.

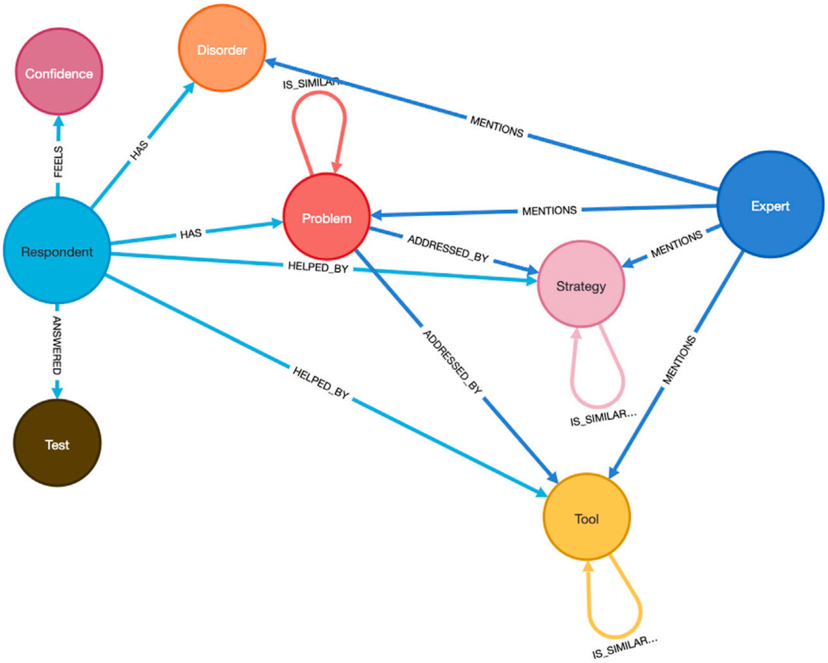


Fig. 6. Final graph representation of the modeled schema.

5. Discussion and results

This section evaluates the performance of the proposed data extraction and integration pipeline. The feasibility of integrating such methods into the overall data engineering pipeline is assessed through the computation of common evaluation metrics. The section concludes with key takeaways and implications from this work.

5.1. Named entity recognition

The NER conducted by the LLM loads a total of 1,011 `PST` and `Disorder` nodes, with the `Strategy` nodes forming the largest group of 360 distinct names (see full breakdown in Appendix Table C.6). From these nodes, 345 relationships were imported into the database (see full breakdown in Appendix Table C.7). Evaluating nodes and relationships generated by the LLM is challenging since the data sources are unstructured. The paragraph containing the exact location of entities is recorded to assess their actual relevance and validity. A common approach to evaluate a model's NER is to compute precision, recall, and the consequent F1-score [6,44,66–68].

A quality evaluation dataset was designed by sampling 10% of the raw chunks from the interview transcripts along with their respective generated nodes and relationships. We concede that such approach can be prone to sampling bias. In fact, studies have considered this to be a demonstrative approach but have also noted the possibility of having a high variance in the results after sampling repetitions [16,69]. Nevertheless, this demonstration could provide an understanding of whether further investigation into using such tools is worthwhile. The sample was stratified such that it represented content from all the experts. Three reviewers were then tasked to collectively read the sampled chunks and perform a manual NER to establish a ground truth of nodes and relationships. Their results were then compared to the ones' extracted from the model by recording the true positives (correctly identified node/relationship), false positives (falsely identified node/relationship), and false negatives (unidentified node/relationship). Table 2 below summarizes the results of the evaluation.

Table 2
Sample quality evaluation of GPT-3.5-Turbo on NER Task (in %).

Entity	Recall	Precision	F1-score
All	75.41	69.84	72.49
Nodes	89.84	75.11	81.80
Relationships	52.87	58.64	55.34

In terms of node extraction, an F1-score of 81.8% is high considering that the LLM has not been fine-tuned on this project's defined node types and rather simply given definitions with two corresponding examples. The reviewers noted that some of the sentences in the chunks were difficult to understand as a result of missing words or incorrect transcribing of speech-to-text. Data quality is surely a limitation that is difficult to improve without introducing extra steps that may limit the scalability of the pipeline. Looking further into the defects, an analysis of the false positives in the sample found that there are a few relevant examples that were assigned to the wrong node type. The `Disorder` nodes were the ones most affected by this issue. However, these nodes may hold a low impact on the overall database, as theoretically their weak semantic similarity to any of the nodes originating from the structured sources would lead them to have a very low graph degree. Other false positives were found to be due to node names composed of one word only, bearing no real meaning as a standalone. One such example of that was "Stubbornness", which was extracted by the LLM to be a `Problem` node. Such naming causes interpretation issues. One could wonder, for instance, if the problem refers to "dyslexics being stubborn", which would be completely wrong and misleading. After tracing back the chunk, "Stubbornness" was actually referring to the "stubbornness of teachers that sometimes refuse to accommodate the learning needs of Dyslexic students". As a consequence, a description attribute was later introduced as a takeaway from this issue: effectively backing up each node name with a contextual and detailed sentence. This description attribute was generated after completing the NER step by feeding the chunks again to the LLM, but this time with the extracted node names as context. In fact, this description attribute was integrated as a way to improve the quality of the Data Resolution task described in Section 5.2.

Since the relationships are extracted directly from the resulting nodes, the F1-score of 55.34% is unsurprisingly lower. An incorrect node classification automatically flags its relationships as false. Other research seems to find similar patterns in performances between nodes and relationships [44]. Therefore, the evaluation of extracted relationships should not be scrutinized with the same breadth. Interestingly, precision, fared higher than recall. This lower recall was amplified by missing nodes from the node extraction task. The phenomenon was found to be especially true when the LLM failed to find `Disorder` nodes, which in turn caused the model to miss relationships with several distinct `Problem` nodes.

Overall, the results are very promising. The automatic pipeline was able to process all the interview transcripts and load the extracted information in about 1 h, which could be made significantly shorter if task parallelization was introduced. In comparison, the human evaluation, comprising of three reviewers, working together to identify all the nodes and relationships for only a 10% sample, took 2.5 h. Based on this, a naive estimate for a fully manual NER could be assumed to be about 25 h. This is excluding the fatigue that could ensue over time and the breaks (in days) which could be required. Thus, the entity extraction by humans could easily take a few days for only 10 interview transcripts. The proposed automatic pipeline therefore surely offers strong potential gains in terms of time consumption.

While the current approach already offers encouraging results, there are several avenues that can be explored to improve the NER task's F1-score. An LLM could be fine-tuned to learn the nodes and relationships in a domain-specific way. This would require creating a ground truth and also accept that the model would be specialized on a certain corpus of nodes and relationships [70]. Another potential solution is retrieval augmented generation (RAG) [70], a method that enriches the context provided to the LLM using a knowledge base. In fact, RAG can be further enhanced by following a framework [71]. Finally, the prompt generation can be delegated to a secondary LLM that has been fine-tuned to generate instructions specific to NER [72,73]. The benefits of such potential improvements can apply to a wide array of tasks, including the one covered in Section 5.2.

5.2. Data resolution

The data resolution task can be thought of as a binary classification task. Hence, the same metrics of Section 5.1 shall be used. There were 17,981 potential similar pair of nodes extracted from the questionnaires and interviews. In a real-world setting, it would not be practical to evaluate the total set and so 2% of the pairs were evaluated. Two reviewers were tasked to determine whether a pair of nodes were similar by simply labeling 1, when similar, or 0 otherwise. The reviewers were provided the same instructions as the LLM to define the context of when to classify a pair of phrases as similar. They were also provided node descriptions to help understand the context of nodes extracted from the interviews, as described in Section 5.1. Finally, the reviewers were privy to the node type of each assessed pair. This information was not provided to the LLM to prevent entity matching biases, potentially induced by the pair sharing the same node type.

The sampling strategy was meticulously designed to ensure an equal distribution between positive and negative instances to diligently evaluate the data resolution task. As the dataset was significantly unbalanced (thought to have less than 10% of positive examples), a special method was adopted to streamline the sample creation. The 17,981 pairs were sorted in descending order of pairwise cosine similarity to increase the likelihood of sampling positive examples. The group of reviewers consequently determined whether a pair was similar until 1% of the total number of pairs was filled with positive examples. Evidently, as a result of the previously mentioned imbalance in the dataset, an equal number of negative examples were also identified through this iterative procedure. The fact that all these negative examples were sourced from the pool of high cosine similarity indicates that it is more challenging for the LLM to avoid false positives compared to resorting to random sampling.

Table 3 outlines the results of the LLM on the data resolution task for the selected sample. The results of the LLM were benchmarked against a baseline model that clustered the node names' textual embeddings using OpenAI's "ada-002" model. This baseline approach aims at grouping similar nodes together. It effectively identifies synonymous entities originating from different data sources, categorizing them under common cluster identifiers. The clustering was conducted using k-means, assigning the optimal value of k based on the highest average silhouette score.

Table 3
Results of the data resolution task (in %).

Model	Node type	Precision	Recall	F1-score
Baseline: Clustered Embeddings	Problem	**91.67**	29.72	44.40
	Tool	36.00	24.32	23.03
	Strategy	**85.71**	16.22	27.27
	Total	59.09	23.42	33.54
GPT-3.5-Turbo	Problem	63.83	**81.08**	**71.42**
	Tool	**63.33**	**51.35**	**56.72**
	Strategy	80.77	**56.76**	**66.67**
	Total	**67.96**	**63.06**	**65.42**

The LLM outperformed the baseline on all metrics when looking simply at the "Total" values, achieving a final F1-score of 65.42%. The Baseline outperformed only on the precision metric of the `Problem` and `Strategy` nodes. Relying on textual embeddings, the Baseline model reached high precision by simply finding most of the syntactic similarities such as "Reading Difficulties" and "Difficulty to Read". However, as illustrated by its poor recall, this model is unable to satisfy the requirements for contextual and thematic similarities such as between "Reading Difficulties" and "Size of Text" or between "Text with every other line highlighted" and "use colors to underline text". This is interesting considering that we expected that the LLM would be at a disadvantage as a result of our sampling strategy biasing toward higher cosine similarity, hypothesized to benefit the clustering of textual embeddings. The higher precision for `Problem` and `Strategy` nodes is therefore explained by the model only classifying a pair as similar in a very small portion of instances, limiting the chances of causing false positives. It is somewhat surprising that the precision of the Baseline on `Tool` was very low. Upon investigation of the examples, it was found that many of the unrelated pairs of `Tool` names contained the word "Dyslexic" or "Dyslexia", increasing their cosine similarity and misleading the baseline model to generate false positives. Considering this, it is impressive that the LLM was able to cope with such pitfalls and classify correctly such nuanced examples as those provided above. It is worth noting however that in a considerable number of false positive examples, the model was providing too broad justifications for thematic similarity. For example, a pair of `Problem` nodes were labeled as similar because they were both "describing a difficulty in an educational setting" — the definition of the `Problem` node type. Ironically, a prompt optimization that attempted to correct this by giving the model context about the pair's node type yielded a slightly lower precision.

In addition to the potential improvements proposed in Section 5.1, one can simply use a more advanced model like GPT-4, which has been shown to yield higher F1-scores at entity resolution [74]. Moreover, one can change the prompt to only focus on syntactic similarities if one faces a use-case that does not require such implicit definition of similarity for data integration.

However, it could be more interesting to change the conception of the data modeling to accommodate for the fact that language is in reality nuanced and that not all relationships are simply syntactic in nature. For example, the prompt can be modified to also provide a confidence score if a pair is deemed similar [75]. Even though some studies observed that such method yielded case-dependent results [76], this probability could be stored as an edge attribute of the similarity link, allowing for a more in-depth analysis within the graph database. Alternatively, the modeling of the relationship, *IS_SIMILAR*, can be modified to allow for three different possible relationships between two nodes from different data sources. For example, the relationships could be *IS_SYNTACTIC_SIMILAR*, *IS_THEMATIC_SIMILAR*, and *IS_FUNCTION_SIMILAR*: the three possible contextual similarities defined to the LLM, as mentioned in Section 4.4. This further exemplifies the importance of conceptual data modeling. In fact, [77] has constructed a semantic framework to help human experts define a more comprehensive strategy to dealing with similarities when attempting to integrate heterogeneous data sources. Combining such frameworks with our explored methodology may enhance the semantic capabilities of LLMs.

5.3. Key takeaways and implications

To summarize, this research did not aim to find the perfect automatic tool for data integration, but to explore the potential of Large Language Models (LLMs) in enhancing this process. The findings reveal that LLMs hold great promise. Minor adjustments to prompts significantly impacted F1-score (increased by a factor of 1.76 in the case of entity resolution), highlighting the sensitivity of these models. Data modeling proved invaluable for crafting effective prompts and contextualizing the instructions, reinforcing the idea that while technology aids, it cannot replace the foundational task of data modeling. Post-processing the LLM's output emerged as a critical step, addressing issues like formatting errors, token limits, and incorrect node or relationship generation. This underscores the importance of a robust data integration pipeline to manage such challenges, indicating areas for further refinement and exploration in the realm of data privacy and processing efficiency.

Acknowledging the limitations of our work is equally important for a comprehensive understanding. The work has shown that the data engineering pipeline can be automated in a manner to aid humans. However, it is still unclear how such approach can be scaled to big data applications. The computational complexity of the tasks, especially that of entity resolution, could pose a problem in cases of high volumes. Regardless, such methodology can prove to be vital to practitioners not operating in such cases. Another limitation stems from the inherent bias associated with using a sample evaluation. This does not diminish the conclusions themselves, however it is important to work and establish a comprehensive sampling framework to evaluate such large datasets considering that a full evaluation is probably unrealistic in most use-cases. Finally, the data privacy concerns relating to using LLMs cannot be ignored. If such concerns arise, one could rely on open-source models, such as Mixtral [78], Mistral [79] or Llama2 [80], if the right resources are available.

6. Conclusion

This study demonstrates the effectiveness of a novel application of Large Language Models (LLMs) for integrating heterogeneous data sources into a graph database. Through a comprehensive methodology that includes data modeling, extraction, and integration, supported by technologies such as Neo4j and GPT-3.5-Turbo, complex data processing tasks can potentially be streamlined. Although the data modeling choices have been centered around one specific dataset, several steps such as those relating to the modeling of entities as well as the decision of where to store attributes can be expanded to other use-cases. This is especially applicable in the context of an educational environment. The evaluation of both Named Entity Recognition and Data Resolution tasks illustrates the effectiveness and efficiency of LLMs in handling diverse data types. The project highlights the synergy between human expertise in data curation and AI's capabilities: opening avenues for more nuanced and scalable research databases.

Our future work aims to develop a more robust framework for data modeling that can better capture the complexities of educational data. The development of such a framework could also include an exploration to enhance an LLM's understanding of nodes and relationships by leveraging techniques such as retrieval augmented generation (RAG) and further prompt-engineering. In addition, data modeling can be improved by accounting for the nuanced nature of language, potentially employing probabilistic approaches to similarity and exploring the inclusion of syntactic, thematic, and functional relationships into the conceptual schema. Moreover, since model fine-tuning is difficult due the lack of available ground truth, it is worthwhile investigating generating a synthetic dataset using LLMs that are specifically tailored to the use-case [81]. It has also been established that different results could be obtained from repeated executions of LLMs [82]. To assess the robustness of the proposed approach, it could be interesting to perform a statistical analysis on multiple runs of the data extraction and integration processes. This quantitative evaluation could also provide the opportunity to compare the robustness of different language models on this specific task. Finally, the optimization of the disambiguation process presents a rich avenue for further research that is not covered here, as this study primarily focused on data extraction and resolution.

CRediT authorship contribution statement

Adel Remadi: Writing – review & editing, Writing – original draft, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Karim El Hage:** Writing – review & editing, Writing – original draft, Validation, Software, Project administration, Methodology, Investigation, Formal analysis, Conceptualization. **Yasmina Hobeika:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Francesca Bugiotti:** Writing – review & editing, Supervision, Resources, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Francesca Bugiotti reports financial support and equipment, drugs, or supplies were provided by European Union Committee within the Erasmus+ Program. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Acknowledgment

This research was co-funded by European Union Committee within the Erasmus+ Program 2014-2020-Key Action 2: Strategic Partnership Project (Agreement n. 2020-1-IT02-K203-080006). We also thank Ruining Ma and Victor Hong for their contribution in our previous publication.

Appendix A. Nature of data used

See Tables A.4 and A.5.

Appendix B. Sample prompts for data integration

See Figs. B.7 and B.8.

Appendix C. Summary of data extracted from expert interviews using LLM

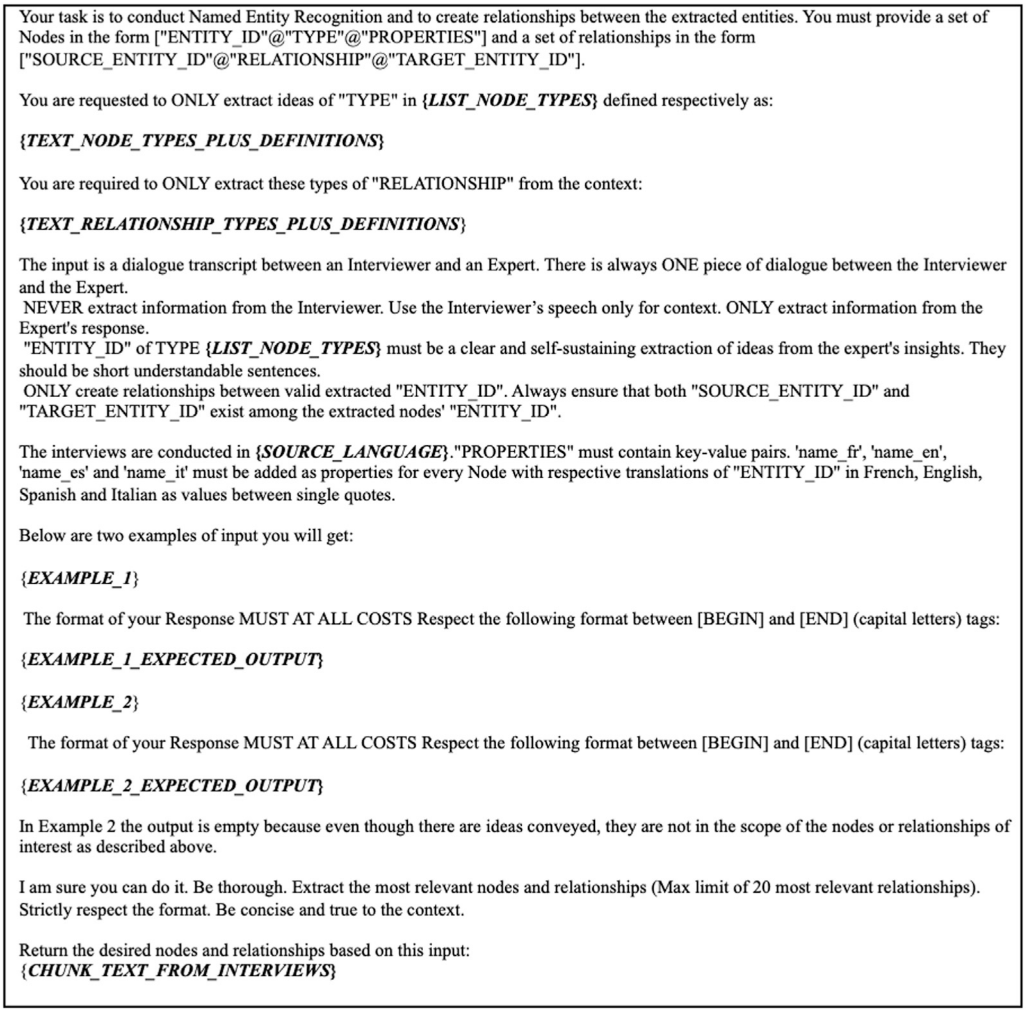
See Tables C.6 and C.7.

Table A.4
Example data from select columns of the questionnaire.

Variable	Respondents	
id	155	34
How old are you ? (do not enter your date of birth)	22	229
Gender	F	Prefer Not To Say
Dyslexics in Family	Mother, Brother	-
What university are you from?	Nanterre Univ.	CentraleSupélec
Are you dyslexic?	Yes	No
Have you been diagnosed with dyslexia?	Yes	-
IF YOU ANSWER YES TO THE PREVIOUS QUESTION - What other difficulty(s) do you have besides dyslexia? [Calculation difficulty - dyscalculia]	Yes	-
IF YOU ANSWER YES TO THE PREVIOUS QUESTION - What other difficulty(s) do you have besides dyslexia? [Other]	-	-
Reading Difficulties	5	2
Presentation Attention	4	4
Audiobook Quality	I don't know	2
Images for Words	4	2
Oral Exams	3	1

Table A.5
Example data from select columns of the VR set.

Variable	Respondents	
id	361	362
created_at	2022-12-12 10:56	2022-12-12 18:00
age	22	32
sex	female	male
dyslexia_type	Dysorthography	Dyscalculia
language	4	4
"Press quickly and twice in a row the yellow button" Time	81.0019	44.9134
"Press quickly and twice in a row the yellow button" Correct	TRUE	TRUE
"Try to say the word kiss/bisous/beso/bacio" Time	0	64.3253
"Try to say the word kiss/bisous/beso/bacio" Correct	FALSE	TRUE
"I feel that I am a person of worth, at least on an equal plane with others"	1	2
"I feel that I have a number of good qualities"	1	2



Your task is to conduct Named Entity Recognition and to create relationships between the extracted entities. You must provide a set of Nodes in the form ["ENTITY_ID"@"TYPE"@"PROPERTIES"] and a set of relationships in the form ["SOURCE_ENTITY_ID"@"RELATIONSHIP"@"TARGET_ENTITY_ID"].

You are requested to ONLY extract ideas of "TYPE" in ***{LIST_NODE_TYPES}*** defined respectively as:

{TEXT_NODE_TYPES_PLUS_DEFINITIONS}

You are required to ONLY extract these types of "RELATIONSHIP" from the context:

{***TEXT_RELATIONSHIP_TYPES_PLUS_DEFINITIONS***}

The input is a dialogue transcript between an Interviewer and an Expert. There is always ONE piece of dialogue between the Interviewer and the Expert.

NEVER extract information from the Interviewer. Use the Interviewer's speech only for context. ONLY extract information from the Expert's response.

"ENTITY_ID" of TYPE ***{LIST_NODE_TYPES}*** must be a clear and self-sustaining extraction of ideas from the expert's insights. They should be short understandable sentences.

ONLY create relationships between valid extracted "ENTITY_ID". Always ensure that both "SOURCE_ENTITY_ID" and "TARGET_ENTITY_ID" exist among the extracted nodes' "ENTITY_ID".

The interviews are conducted in ***{SOURCE_LANGUAGE}***."PROPERTIES" must contain key-value pairs. 'name_fr', 'name_en', 'name_es' and 'name_it' must be added as properties for every Node with respective translations of "ENTITY_ID" in French, English, Spanish and Italian as values between single quotes.

Below are two examples of input you will get:

{***EXAMPLE_1***}

The format of your Response MUST AT ALL COSTS Respect the following format between [BEGIN] and [END] (capital letters) tags:

{***EXAMPLE_1_EXPECTED_OUTPUT}***

{***EXAMPLE_2***}

The format of your Response MUST AT ALL COSTS Respect the following format between [BEGIN] and [END] (capital letters) tags:

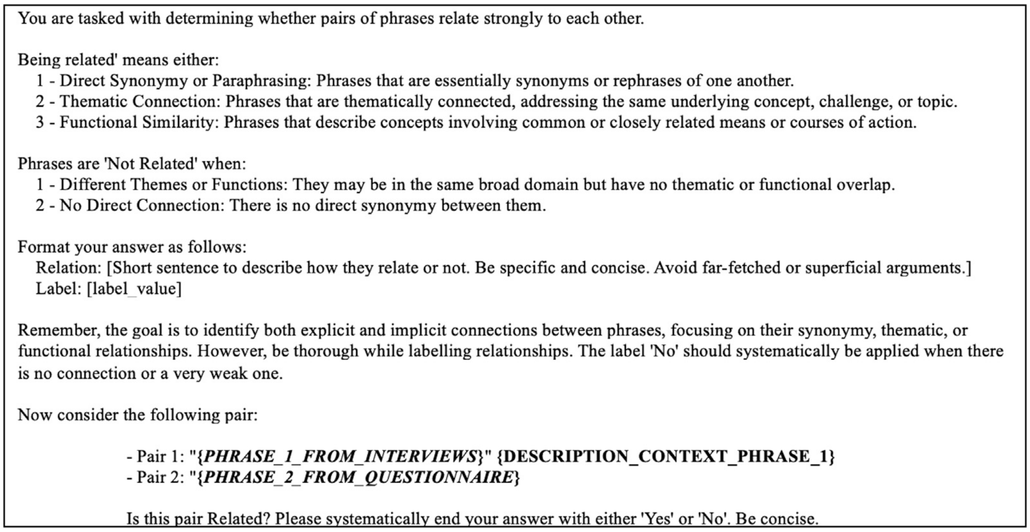
{***EXAMPLE_2_EXPECTED_OUTPUT}***

In Example 2 the output is empty because even though there are ideas conveyed, they are not in the scope of the nodes or relationships of interest as described above.

I am sure you can do it. Be thorough. Extract the most relevant nodes and relationships (Max limit of 20 most relevant relationships). Strictly respect the format. Be concise and true to the context.

Return the desired nodes and relationships based on this input:
{***CHUNK_TEXT_FROM_INTERVIEWS}***

Fig. B.7. Excerpt from prompt used for NER task (text formatted as ***TEXT*** are user inputs).



You are tasked with determining whether pairs of phrases relate strongly to each other.

Being related' means either:
1 - Direct Synonymy or Paraphrasing: Phrases that are essentially synonyms or rephrases of one another.
2 - Thematic Connection: Phrases that are thematically connected, addressing the same underlying concept, challenge, or topic.
3 - Functional Similarity: Phrases that describe concepts involving common or closely related means or courses of action.

Phrases are 'Not Related' when:
1 - Different Themes or Functions: They may be in the same broad domain but have no thematic or functional overlap.
2 - No Direct Connection: There is no direct synonymy between them.

Format your answer as follows:
Relation: [Short sentence to describe how they relate or not. Be specific and concise. Avoid far-fetched or superficial arguments.]
Label: [label_value]

Remember, the goal is to identify both explicit and implicit connections between phrases, focusing on their synonymy, thematic, or functional relationships. However, be thorough while labelling relationships. The label 'No' should systematically be applied when there is no connection or a very weak one.

Now consider the following pair:

- Pair 1: "***{PHRASE_1_FROM_INTERVIEWS}***" **{DESCRIPTION_CONTEXT_PHRASE_1}**
- Pair 2: "***{PHRASE_2_FROM_QUESTIONNAIRE}***

Is this pair Related? Please systematically end your answer with either 'Yes' or 'No'. Be concise.

Fig. B.8. Excerpt from prompt used for Entity Resolution task (text formatted as ***TEXT*** are user inputs).

Table C.6
Extracted nodes from the interview transcripts using LLM.

Entity	Number of nodes
Disorder	61
Problem	314
Tool	276
Strategy	360

Table C.7
Extracted relationships from the interview transcripts using LLM.

Entity pair	Number of relationships
(Problem, Disorder)	83
(Problem, Tool)	125
(Problem, Strategy)	137

References

[1] A. Doan, A.Y. Halevy, Z.G. Ives, Principles of Data Integration, Morgan Kaufmann, 2012.

[2] K. El Hage, A. Remadi, Y. Hobeika, R. Ma, V. Hong, F. Bugiotti, A multi-source graph database to showcase a recommender system for dyslexic students, in: IEEE International Conference on Big Data, BigData, 2023, pp. 3134–3138, http://dx.doi.org/10.1109/BigData59044.2023.10386535.

[3] Y. Tang, X. Wu, C. Zhou, G. Zhu, J. Song, G. Liu, Z. Li, Automatic schema construction of electrical graph data platform based on multi-source relational data models, Data Knowl. Eng. 145 (2023) 761–765, http://dx.doi.org/10.1016/j.datak.2022.102129.

[4] M. Barbella, G. Tortora, A semi-automatic data integration process of heterogeneous databases, Pattern Recognit. Lett. 166 (C) (2023) 134–142, http://dx.doi.org/10.1016/j.patrec.2023.01.007.

[5] Y. Li, J. Li, Y. Suhara, A. Doan, W.-C. Tan, Deep entity matching with pre-trained language models, Proc. VLDB Endow. 14 (1) (2020) 50–60, http://dx.doi.org/10.14778/3421424.3421431.

[6] P. Li, P. Sun, Q. Tang, H. Yan, Y. Wu, X. Huang, X. Qiu, CodeIE: Large code generation models are better few-shot information extractors, in: Annual Meeting of the Association for Computational Linguistics, ACL, 2023, pp. 15339–15353, http://dx.doi.org/10.18653/v1/2023.acl-long.855.

[7] S. Wang, X. Sun, X. Li, R. Ouyang, F. Wu, T. Zhang, J. Li, G. Wang, Gpt-NER: Named entity recognition via large language models, 2023, http://dx.doi.org/10.48550/arXiv.2304.10428, arXiv preprint.

[8] A. Halevy, J. Dwivedi-Yu, Learnings from data integration for augmented language models, 2023, http://dx.doi.org/10.48550/arXiv.2304.04576, arXiv preprint.

[9] R.C. Fernandez, A.J. Elmore, M.J. Franklin, S. Krishnan, C. Tan, How large language models will disrupt data management, Proc. VLDB Endow. 16 (11) (2023) 3302–3309, http://dx.doi.org/10.14778/3611479.3611527.

[10] R. Lukyanenko, A. Castellanos, J. Parsons, M.C. Tremblay, V.C. Storey, Using conceptual modeling to support machine learning, in: International Conference on Advanced Information Systems Engineering, Vol. 350, CAiSE, Springer, 2019, pp. 170–181, http://dx.doi.org/10.1007/978-3-030-21297-1_15.

[11] W. Maass, V.C. Storey, Pairing conceptual modeling with machine learning, Data Knowl. Eng. 134 (2021) 101909, http://dx.doi.org/10.1016/J.DATAK.2021.101909.

[12] J. Trujillo, K. Davis, X. Du, E. Damiani, V. Storey, Conceptual modeling in the era of big data and artificial intelligence: Research topics and introduction to the special issue, Data Knowl. Eng. 135 (2021) http://dx.doi.org/10.1016/j.datak.2021.101911.

[13] S. Arora, B. Yang, S. Eyuboglu, A. Narayan, A. Hojel, I. Trummer, C. Ré, Language models enable simple systems for generating structured views of heterogeneous data lakes, Proc. VLDB Endow. 17 (2) (2023) 92–105, http://dx.doi.org/10.14778/3626292.3626294.

[14] Z. Chen, Z. Gu, L. Cao, J. Fan, S. Madden, N. Tang, Symphony: Towards natural language query answering over multi-modal data lakes, in: Conference on Innovative Data Systems Research, CIDR, 2023.

[15] P. Arocena, B. Glavic, R. Ciucanu, R. Miller, The ibench integration metadata generator, Proc. VLDB Endow. 9 (2015) http://dx.doi.org/10.14778/2850583.2850586.

[16] A. Narayan, I. Chami, L. Orr, C. Ré, Can Foundation Models Wrangle Your Data? Proc. VLDB Endow. 16 (4) (2022) 738–746, http://dx.doi.org/10.14778/3574245.3574258.

[17] A. Halevy, Y. Choi, A. Floratou, M.J. Franklin, N. Noy, H. Wang, Will LLMs reshape, supercharge, or kill data science? Proc. VLDB Endow. 16 (12) (2023) 4114–4115, http://dx.doi.org/10.14778/3611540.3611634.

[18] U. Sivarajah, M.M. Kamal, Z. Irani, V. Weerakkody, Critical analysis of big data challenges and analytical methods, J. Bus. Res. 70 (1) (2017) 263–286, http://dx.doi.org/10.1016/j.jbusres.2016.08.001.

[19] K. Sahatqija, J. Ajdari, X. Zenuni, B. Raufi, F. Ismaili, Comparison between relational and NOSQL databases, in: Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2018, Opatija, Croatia, May 21-25, 2018, IEEE, 2018, pp. 216–221, http://dx.doi.org/10.23919/MIPRO.2018.8400041.

[20] R. Angles, C. Gutierrez, Survey of graph database models, ACM Comput. Surv. 40 (1) (2008) 1–39, http://dx.doi.org/10.1145/1322432.1322433.

[21] R. Angles, C. Gutierrez, An introduction to graph data management, in: Graph Data Management, Fundamental Issues and Recent Developments, in: Data-Centric Systems and Applications, Springer, 2018, pp. 1–32, http://dx.doi.org/10.1007/978-3-319-96193-4_1.

[22] F. Schummer, M. Hyba, An approach for system analysis with model-based systems engineering and graph data engineering, Data-Centric Eng. 3 (2022) e33, http://dx.doi.org/10.1017/dce.2022.33.

[23] A. Nayak, A. Poriya, D. Poojary, Type of NOSQL databases and its comparison with relational databases, Int. J. Appl. Inf. Syst. 5 (4) (2013) 16–19, http://dx.doi.org/10.5120/ijais12-450888.

[24] C. Cattuto, M. Quaggiotto, A. Panisson, A. Averbuch, Time-varying social networks in a graph database: a Neo4j use case, in: International Workshop on Graph Data Management Experiences and Systems, GRADES, ACM, 2013, pp. 1–6, http://dx.doi.org/10.1145/2484425.2484442.

[25] P. Atzeni, F. Bugiotti, L. Cabibbo, R. Torlone, Data modeling in the NoSQL world, Comput. Stand. Interfaces 67 (2020) http://dx.doi.org/10.1016/J.CSI.2016.10.003.

[26] M. Hewasinghage, N.B. Seghouani, F. Bugiotti, Modeling strategies for storing data in distributed heterogeneous NoSQL databases, in: International Conference on Conceptual Modeling, Vol. 11157, ER, Springer, 2018, pp. 488–496, http://dx.doi.org/10.1007/978-3-030-00847-5_35.

[27] V.C. Storey, R. Lukyanenko, A. Castellanos, Conceptual modeling: Topics, themes, and technology trends, ACM Comput. Surv. 55 (14s) (2023) http://dx.doi.org/10.1145/3589338.

[28] I. Davies, P. Green, M. Rosemann, M. Indulska, S. Gallo, How do practitioners use conceptual modeling in practice? Data Knowl. Eng. 58 (3) (2006) 358–380, http://dx.doi.org/10.1016/j.datak.2005.07.007.

[29] M.A. Zaidi, Conceptual modeling interacts with machine learning - A systematic literature review, in: Computational Science and Its Applications, Vol. 12957, ICCSA, Springer, 2021, pp. 522–532, http://dx.doi.org/10.1007/978-3-030-87013-3_39.

[30] A. Garmendia, D. Bork, M. Eisenberg, T. do Nascimento Ferreira, M. Kessentini, M. Wimmer, Leveraging artificial intelligence for model-based software analysis and design, in: Optimising the Software Development Process with Artificial Intelligence, Springer, 2023, pp. 93–117, http://dx.doi.org/10.1007/978-981-19-9948-2_4.

[31] W.S. Lim, M. Butrovich, W. Zhang, A. Crotty, L. Ma, P. Xu, J. Gehrke, A. Pavlo, Database gyms, in: Conference on Innovative Data Systems Research, CIDR, 2023.

[32] D. Bork, S.J. Ali, B. Roelens, Conceptual modeling and artificial intelligence: A systematic mapping study, 2023, http://dx.doi.org/10.48550/arXiv.2303.06758, arXiv preprint.

[33] D. Wu, W. Feng, T. Li, Z. Yang, Evaluating the intelligence capability of smart homes: A conceptual modeling approach, Data Knowl. Eng. 148 (2023) 102218, http://dx.doi.org/10.1016/j.datak.2023.102218.

[34] R. Russo, G.D. Giuseppe, A. Vanacore, V.L. Gatta, A. Ferraro, A. Galli, M. Postiglione, V. Moscato, Graph-based approach for European law classification, in: IEEE International Conference on Big Data, BigData, 2023, pp. 1–9, http://dx.doi.org/10.1109/BigData59044.2023.10386684.

[35] N. Nishikawa, S. Fujiwara, Y. Hayamizu, K. Goda, Physical database design for manufacturing business analytics, in: IEEE International Conference on Big Data, BigData, 2023, pp. 1793–1802, http://dx.doi.org/10.1109/BigData59044.2023.10386475.

[36] G. Alonso, N. Ailamaki, S. Krishnamurthy, S. Madden, S. Sivasubramanian, R. Ramakrishnan, Future of database system architectures, in: Companion International Conference on Management of Data, SIGMOD, ACM, 2023, pp. 261–262, http://dx.doi.org/10.1145/3555041.3589360.

[37] A. Kalinowski, D. Datta, Y. An, A scalable approach to aligning natural language and knowledge graph representations: Batched information guided optimal transport, in: IEEE International Conference on Big Data, BigData, 2023, pp. 383–392, http://dx.doi.org/10.1109/BigData59044.2023.10386670.

[38] I. Trummer, DB-BERT: A database tuning tool that "Reads the Manual", in: International Conference on Management of Data, SIGMOD, ACM, 2022, pp. 190–203, http://dx.doi.org/10.1145/3514221.3517843.

[39] A. Gupta, G. Poels, P. Bera, Generating multiple conceptual models from behavior-driven development scenarios, Data Knowl. Eng. 145 (2023) 102141, http://dx.doi.org/10.1016/J.DATAK.2023.102141.

[40] I. Raharjana, D. Siahaan, C. Fatichah, User stories and natural language processing: A systematic literature review, IEEE Access PP (2021) 1, http://dx.doi.org/10.1109/ACCESS.2021.3070606.

[41] M. Kayali, A. Lykov, I. Fountalis, N. Vasiloglou, D. Olteanu, D. Suciu, CHORUS: Foundation Models for Unified Data Discovery and Exploration, 2023, http://dx.doi.org/10.48550/arXiv.2306.09610, arXiv preprint.

[42] M. Urban, D.D. Nguyen, C. Binnig, OmniscientDB: A large language model-augmented DBMS that knows what other DBMSs do not know, in: International Workshop on Exploiting Artificial Intelligence Techniques for Data Management, AiDM, ACM, 2023, http://dx.doi.org/10.1145/3593078.3593933.

[43] I.A.N. Arachchige, L. Ha, R. Mitkov, J.-D. Steinert, Enhancing named entity recognition for holocaust testimonies through pseudo labelling and transformer-based models, in: International Workshop on Historical Document Imaging and Processing, HIP, ACM, 2023, pp. 85–90, http://dx.doi.org/10.1145/3604951.3605514.

[44] S. Carta, A. Giuliani, L. Piano, A.S. Podda, L. Pompianu, S.G. Tiddia, Iterative Zero-Shot LLM Prompting for Knowledge Graph Construction, 2023, http://dx.doi.org/10.48550/arXiv.2307.01128, arXiv preprint.

[45] I. Trummer, From BERT to GPT-3 codex: harnessing the potential of very large language models for data management, Proc. VLDB Endow. 15 (12) (2022) 3770–3773, http://dx.doi.org/10.14778/3554821.3554896.

[46] A. Sharma, X. Li, H. Guan, G. Sun, L. Zhang, L. Wang, K. Wu, L. Cao, E. Zhu, A. Sim, T. Wu, J. Zou, Automatic data transformation using large language model - An experimental study on building energy data, in: IEEE International Conference on Big Data, BigData, 2023, pp. 1824–1834, http://dx.doi.org/10.1109/BigData59044.2023.10386931.

[47] A. Jindal, S. Qiao, S.R. Madhula, K. Raheja, S. Jain, Turning databases into generative AI machines, in: Conference on Innovative Data Systems Research, CIDR, 2024.

[48] P.A. Bernstein, Applying model management to classical meta data problems, in: Conference on Innovative Data Systems Research, CIDR, 2003.

[49] B. Golshan, A. Halevy, G. Mihaila, W.-C. Tan, Data integration: After the teenage years, in: International Conference on Management of Data, SIGMOD, ACM, New York, NY, USA, 2017, pp. 101–106, http://dx.doi.org/10.1145/3034786.3056124.

[50] A.Y. Halevy, A. Rajaraman, J.J. Ordille, Data integration: The teenage years, in: Proc. VLDB Endow., ACM, 2006, pp. 9–16.

[51] Vrailexia, Vrailexia home page, 2023, URL https://vrailexia.eu/.

[52] J. Roitsch, S. Watson, An overview of dyslexia: definition, characteristics, assessment, identification, and intervention, Sci. J. Educ. 7 (2019) 81–86, http://dx.doi.org/10.11648/j.sjedu.20190704.11.

[53] S. Shaywitz, B. Shaywitz, Dyslexia (specific reading disability), Biol. Psychiatry 57 (11) (2005) 1301–1309, http://dx.doi.org/10.1016/j.biopsych.2005.01.043.

[54] VRAILEXIA - Into The Box, 2024, URL https://vrailexia.eu/the-project/io1-in-the-box/.

[55] M. Rosenberg, Society and the Adolescent Self-Image, Princeton University Press, 1965.

[56] Neo4j, Neo4j cypher query language, 2024, URL https://neo4j.com/product/cypher-graph-query-language/.

[57] Neo4j, K-means clustering, 2023, URL https://neo4j.com/docs/graph-data-science/current/algorithms/kmeans/.

[58] Y. Wu, M. Schuster, Z. Chen, Q.V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, J. Klingner, A. Shah, M. Johnson, X. Liu, Ł. Kaiser, S. Gouws, Y. Kato, T. Kudo, H. Kazawa, K. Stevens, G. Kurian, N. Patil, W. Wang, C. Young, J. Smith, J. Riesa, A. Rudnick, O. Vinyals, G. Corrado, M. Hughes, J. Dean, Google's neural machine translation system: Bridging the gap between human and machine translation, 2016, http://dx.doi.org/10.48550/arXiv.1609.08144, arXiv preprint.

[59] M. Johnson, M. Schuster, Q.V. Le, M. Krikun, Y. Wu, Z. Chen, N. Thorat, F. Viégas, M. Wattenberg, G. Corrado, M. Hughes, J. Dean, Google's multilingual neural machine translation system: Enabling zero-shot translation, Trans. Assoc. Comput. Linguist. 5 (2017) 339–351, http://dx.doi.org/10.1162/tacl_a_00065.

[60] A. Ajith, C. Pan, M. Xia, A. Deshpande, K. Narasimhan, InstructEval: Systematic Evaluation of Instruction Selection Methods, 2023, http://dx.doi.org/10.48550/arXiv.2307.00259, arXiv preprint.

[61] S. Min, X. Lyu, A. Holtzman, M. Artetxe, M. Lewis, H. Hajishirzi, L. Zettlemoyer, Rethinking the role of demonstrations: What makes in-context learning work? in: Conference on Empirical Methods in Natural Language Processing, EMNLP, ACL, 2022, pp. 11048–11064, http://dx.doi.org/10.18653/v1/2022.emnlp-main.759.

[62] H. Xu, Y. Sharaf, W. Tan, L. Shen, B. Van Durme, K. Murray, Y. Jin Kim, Contrastive Preference Optimization: Pushing the Boundaries of LLM Performance in Machine Translation, 2024, http://dx.doi.org/10.48550/arXiv.2401.08417, arXiv preprint.

[63] A. Hendy, M. Abdelrehim, A. Sharaf, V. Raunak, M. Gabr, H. Matsushita, Y. Jin Kim, M. Afify, H. Awadalla, How good are GPT models at machine translation? A comprehensive evaluation, 2023, http://dx.doi.org/10.48550/arXiv.2302.09210, arXiv preprint.

[64] R. Cappuzzo, P. Papotti, S. Thirumuruganathan, Creating embeddings of heterogeneous relational datasets for data integration tasks, in: International Conference on Management of Data, SIGMOD, ACM, 2020, pp. 1335–1349, http://dx.doi.org/10.1145/3318464.3389742.

[65] J. Wei, X. Wang, D. Schuurmans, M. Bosma, E.H. hsin Chi, F. Xia, Q. Le, D. Zhou, Chain of thought prompting elicits reasoning in large language models, 2022, http://dx.doi.org/10.48550/arXiv.2201.11903, arXiv preprint, arXiv:2201.11903.

[66] R. Aly, A. Vlachos, R. McDonald, Leveraging type descriptions for zero-shot named entity recognition and classification, in: International Joint Conference on Natural Language Processing, IJCNLP, ACL, 2021, pp. 1516–1528, http://dx.doi.org/10.18653/v1/2021.acl-long.120.

[67] G. Picco, M. Martinez Galindo, A. Purpura, L. Fuchs, V. Lopez, T.L. Hoang, Zshot: An open-source framework for zero-shot named entity recognition and relation extraction, in: Annual Meeting of the Association for Computational Linguistics, ACL, 2023, pp. 357–368, http://dx.doi.org/10.18653/v1/2023.acl-demo.34.

[68] P. Bose, S. Srinivasan, W.C. Sleeman, J. Palta, R. Kapoor, P. Ghosh, A survey on recent named entity recognition and relationship extraction techniques on clinical texts, Appl. Sci. 11 (18) (2021) http://dx.doi.org/10.3390/app11188319.

[69] J. Liu, D. Shen, Y. Zhang, B. Dolan, L. Carin, W. Chen, What makes good in-context examples for GPT-3? in: Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures, ACL, 2022, http://dx.doi.org/10.18653/v1/2022.deelio-1.10.

[70] A. Balaguer, V. Benara, R. Luiz de Freitas Cunha, R. de M. Estevão Filho, T. Hendry, D. Holstein, J. Marsman, N. Mecklenburg, S. Malvar, L. Nunes, R. Padilha, M. Sharp, B. Silva, S. Sharma, V. Aski, R. Chandra, RAG vs Fine-tuning: Pipelines, Tradeoffs, and a Case Study on Agriculture, 2024, http://dx.doi.org/10.48550/arXiv.2401.08406, arXiv preprint.

[71] Q. Sun, K. Huang, X. Yang, R. Tong, K. Zhang, S. Poria, Consistency guided knowledge retrieval and denoising in LLMs for zero-shot document-level relation triplet extraction, 2024, http://dx.doi.org/10.48550/arXiv.2401.13598, arXiv preprint.

[72] N. Mihindukulasooriya, S. Tiwari, C.F. Enguix, K. Lata, Text2KGBench: A benchmark for ontology-driven knowledge graph generation from text, in: The Semantic Web, 2023, pp. 247–265, http://dx.doi.org/10.1007/978-3-031-47243-5_14.

[73] X. Wang, Q. Yang, LingX at ROCLING 2023 multiNER-health task: Intelligent capture of Chinese medical named entities by LLMs, in: Conference on Computational Linguistics and Speech Processing, ACLCLP, 2023, pp. 350–358.

[74] R. Peeters, C. Bizer, Entity matching using large language models, 2023, http://dx.doi.org/10.48550/arXiv.2310.11244, arXiv preprint.

[75] H. Li, L. Feng, S. Li, F. Hao, C.J. Zhang, Y. Song, L. Chen, On leveraging large language models for enhancing entity resolution, 2024, http://dx.doi.org/10.48550/arXiv.2401.03426, arXiv preprint.

[76] N. Nananukul, K. Sisaengsuwanchai, M. Kejriwal, How does prompt engineering affect ChatGPT performance on unsupervised entity resolution? 2023, http://dx.doi.org/10.48550/arXiv.2310.06174, arXiv preprint.

[77] F. Narducci, M. Comerio, C. Batini, M. Castelli, A similarity-based framework for service repository integration, Data Knowl. Eng. 106 (2016) 18–35, http://dx.doi.org/10.1016/j.datak.2016.08.001.

[78] A.Q. Jiang, A. Sablayrolles, A. Roux, A. Mensch, B. Savary, C. Bamford, D.S. Chaplot, D. de las Casas, E.B. Hanna, F. Bressand, G. Lengyel, G. Bour, G. Lample, L.R. Lavaud, L. Saulnier, M.-A. Lachaux, P. Stock, S. Subramanian, S. Yang, S. Antoniak, T.L. Scao, T. Gervet, T. Lavril, T. Wang, T. Lacroix, W.E. Sayed, Mixtral of experts, 2024, http://dx.doi.org/10.48550/arXiv.2401.04088, arXiv preprint.

[79] A.Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D.S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L.R. Lavaud, M.-A. Lachaux, P. Stock, T.L. Scao, T. Lavril, T. Wang, T. Lacroix, W.E. Sayed, Mistral 7B, 2023, http://dx.doi.org/10.48550/arXiv.2310.06825, arXiv preprint.

[80] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C.C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P.S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E.M. Smith, R. Subramanian, X.E. Tan, B. Tang, R. Taylor, A. Williams, J.X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, T. Scialom, Llama 2: Open foundation and fine-tuned chat models, 2023, http://dx.doi.org/10.48550/arXiv.2307.09288, arXiv preprint.

[81] R. Tang, X. Han, X. Jiang, X. Hu, Does synthetic data generation of LLMs help clinical text mining? 2023, http://dx.doi.org/10.48550/arXiv.2303.04360, arXiv preprint.

[82] E.M. Bender, T. Gebru, A. McMillan-Major, S. Shmitchell, On the dangers of stochastic parrots: Can language models be too big? in: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT '21, Association for Computing Machinery, New York, NY, USA, 2021, pp. 610–623, http://dx.doi.org/10.1145/3442188.3445922.

Adel Remadi holds a Master in Data Science and Business Analytics from CentraleSupélec & ESSEC Business School. Currently, he is an entrepreneur within the Data Science and Tech Space. Prior to that, he garnered extensive experience as a consultant in the Tech industry. His background includes the completion of both an M.Sc. in Management and Economics of Innovation as well as a Master's degree in Industrial Engineering.

Karim El Hage holds a Master in Data Science and Business Analytics from CentraleSupélec & ESSEC Business School. Currently, he is a Data Scientist at Echo Analytics. After completing his Bachelor in Mechanical Engineering from the American University of Beirut, he developed cross-jurisdictional expertise as an Engineer, working with different technologies and applications.

Yasmina Hobeika holds a Master in Data Science and Business Analytics from CentraleSupélec & ESSEC Business School. Currently, she is working as a Business Intelligence Engineer at Amazon, specializing in Historical automation and post processing tasks. Prior to that, she completed a Bachelor in Computer & Communications Engineering at the American University of Beirut.

Francesca Bugiotti holds a position as assistant professor at CentraleSupélec in Paris. She received her "Dr. Ing." degree in Computer Engineering in 2012, with a thesis on heterogeneity in databases. She worked as an intern and as a post-doc at Inria Saclay studying the problem of indexing RDF datasets in a cloud infrastructure and studying efficient data storage mechanisms for heterogeneous data in the cloud and joined CentraleSupélec in 2015. Her research activity focuses on heterogeneous data integration, big data, conceptual models, NoSQL storage systems integration, NoSQL data model characteristics and query expressive power.