Ad-hoc v.s LLM based System for Information Retrieval in Large Tabular Data: A Comparative Study in Public Medicine Procurement Audits

Arthur L. Silva
Federal Rural University of Pernambuco
Recife, Brazil
als.arthurlimasilva@gmail.com

Adriano M. A. Lima
Court of Accounts of Pernambuco
Recife, Brazil
adrianomarabuco@tcepe.tc.br

George Valença
Federal Rural University of Pernambuco
Recife, Brazil
george.valenca@ufrpe.br

George G. Cabral
Federal Rural University of Pernambuco
Recife, Brazil
george.gcabral@ufrpe.br

Abstract

Background: Auditing is key when dealing with public expenses. Despite its importance, frequently auditing efforts must prioritize few targets due to a lack of human resources. However, leveraging the auditing process by developing a system that can automatically process large documents is a feasible task.
Problem: The Information Retrieval (IR) problem considered in this work relies on two components: (i) the text to be searched and (ii) the data source where the required information is supposed to be. The first component is not standardized, presenting a challenge to an automated solution. The second component is structured; however, it is available in a large data source, which may consist of an obstacle for some automated IR methods. Specifically, given a drug specification, our system must find all available products that match this description in a large data source.
Solution: This work investigates two different information retrieval solutions. The first approach basically relies on *apriori* knowledge of the problem for preprocessing the text and computing words similarity. The second approach leverages a powerful LLM to search in the same data source.
IS Theory: Information Processing Theory
Research Method: Proof of Concept
Experimental Results: The results show that the proposed Ad-hoc method reaches accuracies from 72.4% up to 86.9% while the LLM based approach struggles to find satisfactory results mainly by its non-deterministic behavior and the hallucination problem.
Contribution: With regard to the industry, the developed system has the potential to significantly improve the quality and scale of auditing processes. For the academy, the present work unveils limitations of using LLM based approaches for searching in large structured tabular data (± 25000 rows).

CCS Concepts

- **Information systems** → **Environment-specific retrieval**.

Keywords

Information Retrieval, Tabular data, Natural Language Processing, LLM, Audit, Public Procurement, Medicine

1 Introduction

Many auditing problems are quite specific and inherently human tasks, also, in many cases, it can extremely benefit from computational tools [4, 11, 22, 24]. Despite their high level of specific characteristics, a common aspect for this task domain is the need of analyzing a large amount of data. Public expenses consist of a wide range of different classes of products, such as medical equipment, food supply, flight tickets, etc.. Usually, such expenses can only be performed through a formal procedure where a document (public procurement)[1] containing the desired items is publicly released so that a fair bidding procedure can happen. This is the case of the medicines purchasing by public agents.

A public procurement for medicines purchasing procedure usually contains hundreds of different medicines descriptions. A human auditor needs to analyze whether each description is correct and, more than that, needs to check whether the proposed price of the medicine is suitable. This is an exceedingly time consuming task since each medicine description may be produced by a number of different laboratories with different prices. I.e., for each medicine the auditor needs to know all corresponding products available in the market. With the unique identifiers of these products (i.e., a bar code or EAN) at hand, the auditor could, for example, automatically search past purchases of the same medicine to statistically identify whether the proposed price in the public notice is valid or not, for example. Therefore, the main **motivation** of this work is to build and validate a system to automate the process of finding all products publicly available at the medicines' market that match a given description. Automating this task may result in a substantial increase of auditing capabilities.

Traditional Machine Learning classifiers are not applicable to the aforementioned problem since this is not a conventional classification task. On the other hand, Open Domain Question

[1] In this work, the terms public procurement and public notice can be intercambiated employed.

Answering (ODQA) [14, 15] is a research area that already deals with the problem of information retrieval in tabular data [17, 25]. Nonetheless, information retrieval through open-domain question answering (ODQA) using large tables using large language models (LLMs) remains an open challenge due to significant issues such as limited context length, the risk of hallucinations, and the complexities of prompt engineering.

This work develops and compares two systems for information retrieval in large tabular data considering items on public procurements for medicines purchasing. The first system consists of a LLM based query using the version (GPT 4o mini) of the ChatGPT [21]. In addition, we introduce an ad-hoc method, referred to as IR-Med, that operates by preprocessing both, the input data and the tabular data, in order to leverage the search results. The outcome of the provided solution can drastically reduce the human labor in sharply finding the corresponding medicines in the market and consequently scaling up the auditing capability of a human auditor.

In short, the main contributions of this work are:

- Development of an ad-hoc method to search available medicines in the market given ill-defined medicines' descriptions;
- Development of an LLM based method to search available medicines in the market given ill-defined medicines' descriptions; and
- Comparing the performance of the proposed methods on ten public procurements for medicines purchasing.

On top of the above mentioned contributions, the Information Systems area may benefit from this research in the sense that the developed tools will be made open-source and can be accessed by any person interested in validating public medicines purchasing processes. Since documents of public procurements are easy to access, anyone will be able to perform this part of the audit process. More specifically, with regard to the Information Processing Systems scope, this work analyses two different information retrieval approaches working as short-term memory systems.

The remainder of the work is structured as follows: Section 2 presents some related works; Section 3 thoroughly describes the problem; Section 4 introduces the proposed approaches; Section 5 presents the experiments and results discussion; Section 6 presents the treats to validity; and finally, Section 7 brings the conclusions and future works.

2 Related Works

Recent works tackled the problem of processing data from public procurement automatically [7, 22, 26]. Velasco et al. [26] present a decision support system (DSS) developed to detect fraud in public procurement processes in Brazil. The system was created to overcome limitations faced by law enforcement agencies, which traditionally rely on complaints to investigate fraud. The DSS uses data mining algorithms and data science tools to identify patterns of corruption risk, such as collusion between companies, conflicts of interest and shell companies. The application of the DSS in states such as São Paulo and Paraíba resulted in the identification of billions in suspicious contracts and contributed to relevant anti-corruption operations.

Brandão et al. [7] introduce the PLUS system, a semi-automated pipeline designed to detect fraud in public procurement. The proposal addresses the challenges faced when dealing with the diversity and lack of standardization of documents related to public procurement. PLUS was evaluated using public procurement data in the state of Minas Gerais, Brazil. According to the authors, applications include creating audit trails to detect fraud, such as collusion between companies, and building a reference price database to identify overpricing in public procurement. The work highlights the importance of combining data science technologies with the expertise of analysts to improve fraud detection and increase transparency in public procurement processes.

The quality of the solutions for the Open Domain Question Answering problem have become much more satisfactory with the use of LLMs, however, even before the existence of these models, ODQA for tabular data was already a relevant research field [9, 13, 25]. Herzig et al. propose a novel approach to ODQA focused on tabular data rather than traditional textual content. It introduces a dense table retriever (DTR) model, pre-trained to efficiently identify relevant tables from large corpora and enhanced using hard negative examples to improve retrieval accuracy. The authors demonstrate that leveraging table-specific embeddings and dense retrieval methods optimized for tabular data can substantially improve recall and exact match rates for table-based QA tasks.

In the same direction, Chen et al. [9] integrate structured tabular data with unstructured text. They introduce the OTT-QA dataset, requiring multi-hop reasoning across tables and text, making evidence retrieval complex due to the need for interdependent information from both sources. The authors propose two techniques: a "fusion retriever" that combines table segments and relevant passages into context-rich units, and a "cross-block reader" employing sparse attention to process lengthy, interrelated evidence blocks.

The problem considered in this work is related to the previous works in the sense that it is also a question answering problem, however, our problem is a domain specific problem. Therefore, ODQA techniques can be applied to it but the approaches generated in this work do not generalize to other domains.

Similarly to the present work, extracting information from tabular data is also the topic of recent works and surveys [18, 30]. Liu et al. classified approaches to deal with tabular data into three categories: (1) heuristic methods, that are algorithmic straightforward and don't require much effort in engineering or learning [1, 3]; (2) Feature engineering based, that extract statistical and lexical features to use with machine learning models [16, 20]; and (3) Deep learning based [19, 31].

In the last few years, a high number of solutions for domain specific problems are using Large Language Models - LLMs [8, 28]. Furthermore, specific LLMs are being generated for specific domains such as software defect prediction [29] and software testing [27], for example.

Despite of its weaknesses (such as hallucination), by applying LLMs to auditing, it is possible to automate the screening of large volumes of data, identifying patterns and anomalies with greater accuracy and efficiency than traditional methods [12]. These models can process financial documents, identify discrepancies and even predict areas of risk based on historical trends [2]. In addition, LLMs can assist in regulatory analysis by supporting the verification of adherence to standards and policies, and contribute to a more agile audit with a reduced margin of error.

Finally, since LLMs are often used as ODQA models and given the domain specific nature of our problem, this work develops and evaluates an ODQA approach and an heuristic approach for, given a medicine description, finding corresponding products in a large amount of tabular data.

3 Problem Description

Brazilian municipalities produce and release public procurements in order to purchase medicines for public hospitals and other public health services. Nonetheless, all public entities, as a rule, must acquire goods and services by means of a public bid proceeding that ensures equal conditions for all bidders[2].

Definition 1 (Active Pharmaceutical Ingredient) - a.k.a., API, are the raw material to produce medicines. It is the main substance in a medicine and gives its pharmaceutical characteristic. Nonetheless, many medicines are produced as a merging of different active ingredients and in different dosages. These cases pose a challenge to an automatic information retrieval method.

Definition 2 (Pharmaceutical Form) - the pharmaceutical form consists of the form a medicine is presented, e.g. tablet, capsule, solution for injection, cream, etc.

Predominantly, a public procurement for medicines' purchasing contains a long list of medicines descriptions to be acquired. Each item of this list must be specified such that the bidders can undoubtedly identify the item. Nonetheless, this specification must not trace the item to an unique supplier company. For example, the item acetaminophen is available through a variety of different brands, dosages (325 mg, 500 mg, etc.) and pharmaceutical forms (tablet, chewable tablet, liquid oral, etc.). So, the active ingredient (e.g., acetaminophen), dosage, pharmaceutical form and any other information to distinguish the item from other similar medicine must be present without citing any supplier brand or company. This is a rule that can be disrespected in few cases, however, this is not in the scope of this work.

Producing the list of medicines descriptions contained in the public procurement is a human task. In addition, the way how the item is specified in the document does not follow a rigid standardized procedure. Therefore, the author of the public notice can split the information in many columns in a table, or worse, omit information. Auditing a public notice document consists, among other things, in evaluating whether or not each item is satisfactorily described. Unfortunately, frequently the auditing procedure is not conducted by an experienced medicine practitioner or a pharmacist. It is important to emphasize that each public notice contains hundreds of items to be purchased and the amount of notices to be audited, per state of the country, is often incompatible with the number of human auditors.

Given the problem of identifying a proper medicine description for each item in the public notice, a machine learning (or data science) practitioner may be promptly tempted to model a conventional classifier (e.g., a deep learning network or a random forest) in order to, given an input (the medicine description in the public notice) to produce an output containing the proper medicine description. This approach contains some challenges: (i) the lack of an annotated corpus; (ii) the extremely high number of potential classes (if we think in the number of active ingredients as the number of classes, this number is currently 2072 items); (iii) frequently an item in the public notice consists in a combination of many active ingredients, as aforementioned (i.e., in this case, the way the medicine is described hugely affects the classifier output); and (iv) other information such as dosage and form are often not standardized.

Figure 1 depicts some examples of how the medicines' items are described in a public notice (*in Portuguese*). Notice that there isn't a pattern on the information in each table cell. This challenges a suitable identification of a set of corresponding available products that match the description. In addition, item 2 of public notice 2 shows a case where a medicine is formed by more than one active ingredients.

3.1 The CMED list of available medicines

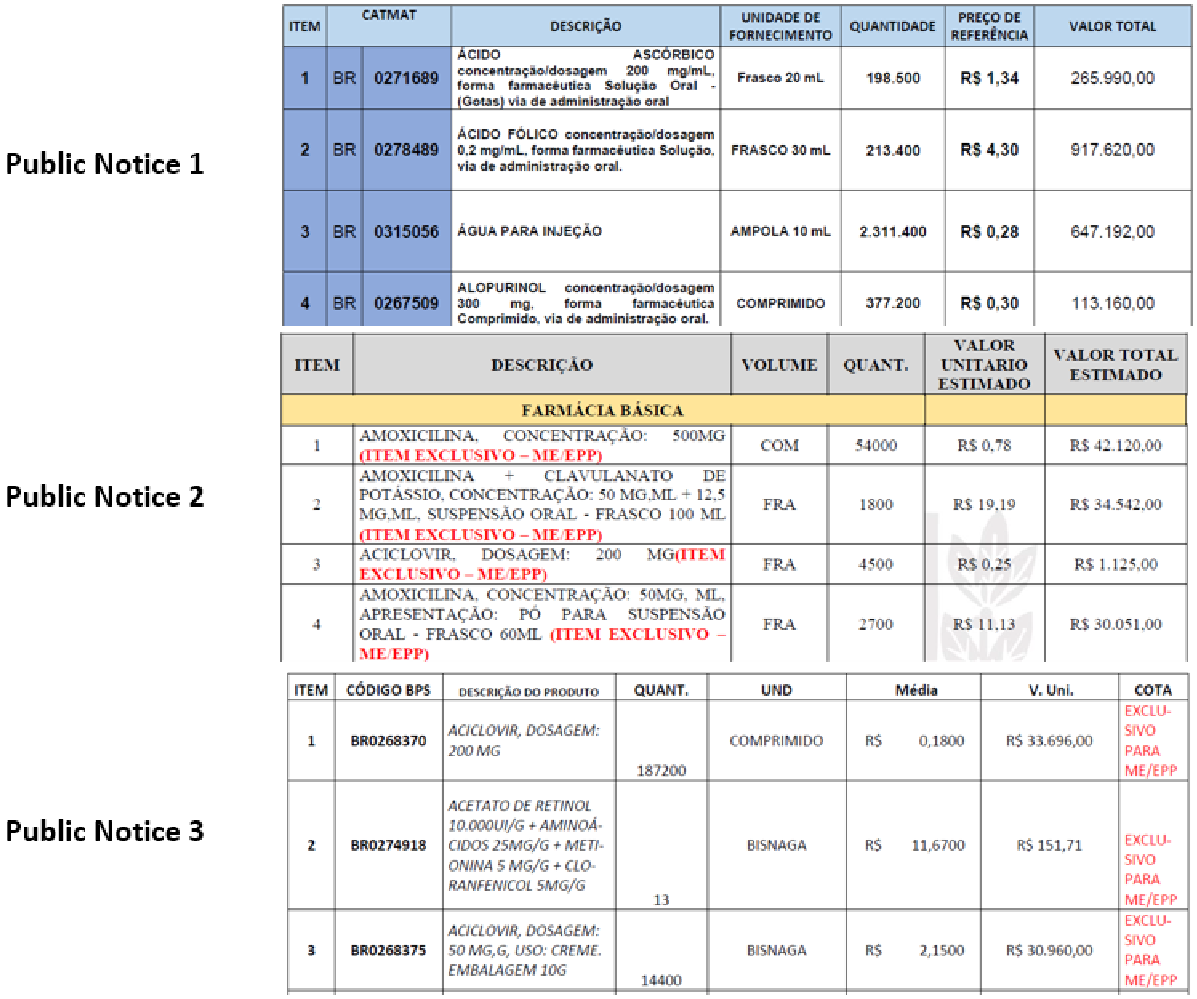
The CMED[3] (*Câmara de Regulação do Mercado de Medicamentos*) is a workgroup under the supervision of the national agency ANVISA (*Agência Nacional de Vigilância Sanitária*). Among others, this group is responsible by the inspection of the prices of medicines in Brazil. All medicines, represented by their structured information, available at the market are cataloged in a monthly updated report. This report releases a table where each row is formed by columns containing information such as: active ingredient; dosage; pharmaceutical form; bar code (i.e., each product available at the market, even similar medicines produced by different companies, have an unique bar code); name of the supplier company; etc. Figure 2 shows a range of rows of the CMED list. In these rows it is possible to find medicines having the same specifications (i.e., *substância* and *apresentação*) but produced by different laboratories.

4 Scientific Methodology

First of all, the present study is aimed at developing a Proof of Concept (PoC), i.e., to demonstrate the feasibility of the provided solution for the following problem:

[2] https://www.lexology.com/library/detail.aspx?g=6f266055-86ba-4849-8dd1-85e07e85b397

[3] https://www.gov.br/anvisa/pt-br/assuntos/medicamentos/cmed/precos



Public Notice 1

ITEM	CATMAT		DESCRIÇÃO	UNIDADE DE FORNECIMENTO	QUANTIDADE	PREÇO DE REFERÊNCIA	VALOR TOTAL
1	BR	0271689	ÁCIDO ASCÓRBICO concentração/dosagem 200 mg/mL, forma farmacêutica Solução Oral - (Gotas) via de administração oral	Frasco 20 mL	198.500	R$ 1,34	265.990,00
2	BR	0278489	ÁCIDO FÓLICO concentração/dosagem 0,2 mg/mL, forma farmacêutica Solução, via de administração oral.	FRASCO 30 mL	213.400	R$ 4,30	917.620,00
3	BR	0315056	ÁGUA PARA INJEÇÃO	AMPOLA 10 mL	2.311.400	R$ 0,28	647.192,00
4	BR	0267509	ALOPURINOL concentração/dosagem 300 mg, forma farmacêutica Comprimido, via de administração oral.	COMPRIMIDO	377.200	R$ 0,30	113.160,00

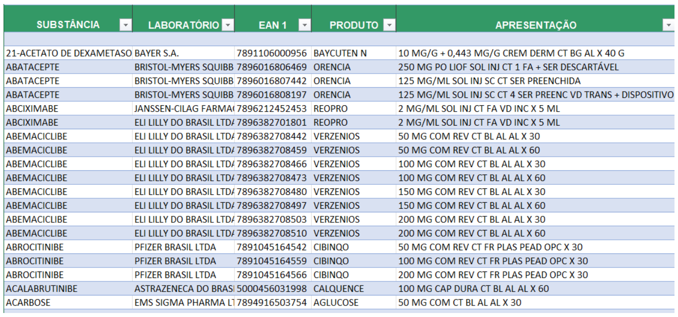
Public Notice 2

ITEM	DESCRIÇÃO	VOLUME	QUANT.	VALOR UNITARIO ESTIMADO	VALOR TOTAL ESTIMADO
FARMÁCIA BÁSICA					
1	AMOXICILINA, CONCENTRAÇÃO: 500MG (ITEM EXCLUSIVO – ME/EPP)	COM	54000	R$ 0,78	R$ 42.120,00
2	AMOXICILINA + CLAVULANATO DE POTÁSSIO, CONCENTRAÇÃO: 50 MG,ML + 12,5 MG,ML, SUSPENSÃO ORAL - FRASCO 100 ML (ITEM EXCLUSIVO – ME/EPP)	FRA	1800	R$ 19,19	R$ 34.542,00
3	ACICLOVIR, DOSAGEM: 200 MG(ITEM EXCLUSIVO – ME/EPP)	FRA	4500	R$ 0,25	R$ 1.125,00
4	AMOXICILINA, CONCENTRAÇÃO: 50MG, ML, APRESENTAÇÃO: PÓ PARA SUSPENSÃO ORAL - FRASCO 60ML (ITEM EXCLUSIVO – ME/EPP)	FRA	2700	R$ 11,13	R$ 30.051,00

Public Notice 3

ITEM	CÓDIGO BPS	DESCRIÇÃO DO PRODUTO	QUANT.	UND	Média	V. Uni.	COTA
1	BR0268370	ACICLOVIR, DOSAGEM: 200 MG	187200	COMPRIMIDO	R$ 0,1800	R$ 33.696,00	EXCLUSIVO PARA ME/EPP
2	BR0274918	ACETATO DE RETINOL 10.000UI/G + AMINOÁCIDOS 25MG/G + METIONINA 5 MG/G + CLORANFENICOL 5MG/G	13	BISNAGA	R$ 11,6700	R$ 151,71	EXCLUSIVO PARA ME/EPP
3	BR0268375	ACICLOVIR, DOSAGEM: 50 MG,G, USO: CREME. EMBALAGEM 10G	14400	BISNAGA	R$ 2,1500	R$ 30.960,00	EXCLUSIVO PARA ME/EPP

Figure 1: Examples of how medicines' items are displayed in public procurements.



SUBSTÂNCIA	LABORATÓRIO	EAN 1	PRODUTO	APRESENTAÇÃO
21-ACETATO DE DEXAMETASO	BAYER S.A.	7891106000956	BAYCUTEN N	10 MG/G + 0,443 MG/G CREM DERM CT BG AL X 40 G
ABATACEPTE	BRISTOL-MYERS SQUIBB	7896016806469	ORENCIA	250 MG PO LIOF SOL INJ CT 1 FA + SER DESCARTÁVEL
ABATACEPTE	BRISTOL-MYERS SQUIBB	7896016807442	ORENCIA	125 MG/ML SOL INJ SC CT SER PREENCHIDA
ABATACEPTE	BRISTOL-MYERS SQUIBB	7896016808197	ORENCIA	125 MG/ML SOL INJ SC CT 4 SER PREENC VD TRANS + DISPOSITIVO
ABCIXIMABE	JANSSEN-CILAG FARMA(	7896212452453	REOPRO	2 MG/ML SOL INJ CT FA VD INC X 5 ML
ABCIXIMABE	ELI LILLY DO BRASIL LTD/	7896382701801	REOPRO	2 MG/ML SOL INJ CT FA VD INC X 5 ML
ABEMACICLIBE	ELI LILLY DO BRASIL LTD/	7896382708442	VERZENIOS	50 MG COM REV CT BL AL AL X 30
ABEMACICLIBE	ELI LILLY DO BRASIL LTD/	7896382708459	VERZENIOS	50 MG COM REV CT BL AL AL X 60
ABEMACICLIBE	ELI LILLY DO BRASIL LTD/	7896382708466	VERZENIOS	100 MG COM REV CT BL AL AL X 30
ABEMACICLIBE	ELI LILLY DO BRASIL LTD/	7896382708473	VERZENIOS	100 MG COM REV CT BL AL AL X 60
ABEMACICLIBE	ELI LILLY DO BRASIL LTD/	7896382708480	VERZENIOS	150 MG COM REV CT BL AL AL X 30
ABEMACICLIBE	ELI LILLY DO BRASIL LTD/	7896382708497	VERZENIOS	150 MG COM REV CT BL AL AL X 60
ABEMACICLIBE	ELI LILLY DO BRASIL LTD/	7896382708503	VERZENIOS	200 MG COM REV CT BL AL AL X 30
ABEMACICLIBE	ELI LILLY DO BRASIL LTD/	7896382708510	VERZENIOS	200 MG COM REV CT BL AL AL X 60
ABROCITINIBE	PFIZER BRASIL LTDA	7891045164542	CIBINQO	50 MG COM REV CT FR PLAS PEAD OPC X 30
ABROCITINIBE	PFIZER BRASIL LTDA	7891045164559	CIBINQO	100 MG COM REV CT FR PLAS PEAD OPC X 30
ABROCITINIBE	PFIZER BRASIL LTDA	7891045164566	CIBINQO	200 MG COM REV CT FR PLAS PEAD OPC X 30
ACALABRUTINIBE	ASTRAZENECA DO BRASI	5000456031998	CALQUENCE	100 MG CAP DURA CT BL AL AL X 60
ACARBOSE	EMS SIGMA PHARMA LT	7894916503754	AGLUCOSE	50 MG COM CT BL AL AL X 30

Figure 2: Part of the CMED list of medicines available in the Brazilian market.

Problem definition - Given a poorly standardized description of a medicine in a public procurement, return all the rows and respective EANs (i.e., bar codes) in the CMED table corresponding to this description.

With the aforementioned aim, two approaches were developed and referred to as IR-Med and ChatGPT-4o Assistant, respectively.

4.1 IR-Med - Modeling Phase

The modeling phase consists in: (i) the pre-processing of the CMED columns *substância* (active ingredients) and *apresentação* (a column comprising the remainder information of a medicine, i.e., form, dosage, etc.); (ii) clustering CMED rows belonging to the same active ingredient; and (iii) relevant tokens extraction of the CMED columns.

4.1.1 Words pre-processing. This phase handles data from both columns of the CMED list. The following treatments are performed: (i) converting the words to lowercase; (ii) removal of the accents; (iii) removval of special characters (i.e., a character not in the characters intervals a-z, A-Z, 0-9, including _); (iv) inserting blank space between numbers and words; (v) based on the ANVISA vocabulary [4], abbreviating the forms (e.g., *comprimido* to *com*); (vi) removal of numbers from the active ingredients descriptions; (vii) removal of stopwords; (viii) removal of repeated words; and (ix) removal of ions and associated chemical compounds such as *cloreto*, *permanganato*, *sulfato*, *brometo*, etc (these terms have shown a detrimental effect in the medicine identification).

4.1.2 Clustering CMED items by their Active Ingredients.

Once the initial pre-processing is carried out, the rows are grouped by their active ingredients. In case of more than one active ingredient, the words are sorted based on their lexical order. Table 1 shows an example of active ingredients and their respective of bar codes (i.e., EANs). This information is then stored in a hash table structure.

Table 1: Stored information on the grouped-cmed hash table.

Active Ingredient(s)	Sorted Active Ingredient(s)	List of rows indexes
zidovudina	zidovudina	[29369, 29370, 29371]
zidovudina lamivudina	zidovudina lamivudina	[29372, 29373, 29374, 29375]
zinco	zinco	[411, 27728, 27729, ...]
zinco nitrato nafazolina	nafazolina nitrato zinco	[27742, 27743]

4.1.3 Identifying relevant words.

Once the grouped-cmed hash table is built, two sets are created: (i) **cmed-ai-words** containing all different words in the pre-processed active ingredients column of the CMED table; and (ii) **cmed-pr-words** containing all different words in the presentation column of the CMED table (Fig. 2, column Apresentação). The words in these sets are used to precisely identify the information on the items descriptions in a public notice.

As an example, consider that after pre-processing a item description of a public notice the following sentence is returned: "amoxicilina 500 mg clavulanato potassio 125 com". The terms amoxicilina and potassio will be associated to the active ingredient and 500 mg 125 com will be acknowledged as the medicine form, dosage, etc. (i.e., present in the *apresentação* column of the CMED table).

4.2 Information Retrieval Phase

Given a medicine description in a public notice, the words in the description are pre-processed according to Section 4.1.1 and compared to the words contained in the sets **cmed-ai-words** and **cmed-pr-words** in order to separate the active ingredients from the remaining information. This description is then splited in two sentences: (i) **desc-ai** - information regarding the active ingredient and (ii) **desc-pr** - information regarding the dosage, form, etc.

Next, **desc-ai** is compared to each active ingredient in **grouped-cmed** hash table and the most similar one is retrieved. This similarity is computed according to the Jaro-Wrinkler similarity (Eq. 1).

$$sim_j w_1, w_2 \begin{cases} 0 & \text{if } m = 0 \\ \frac{1}{3}\left(\frac{m}{|w_1|} \quad \frac{m}{|w_2|} \quad \frac{m-t}{m}\right) & \text{otherwise} \end{cases} \tag{1}$$

where m is the number of matching characters, t is the number of transpositions, and $|w_1|$ and $|w_2|$ are, respectively, the size of the words w_1 and w_2.

Finally, the words in **desc-pr** are then compared to the set (R) of rows associated to the previously found active ingredient. For each row in R, the intersection between the set of words in the column *apresentação* and the set of words in **desc-pr** is computed and the rows in R with higher intersection are then retrieved. At this stage, returning no rows from the CMED list is a high evidence of poorly written description in the public procurement.

A replication package is available at:
https://github.com/ArthurLimaS/ir-med

4.3 ChatGPT-4o Assistant

An assistant model was built to operate specifically on data from the CMED table. This is an example of application of the well-known Retrieval-Augmented Generation (RAG) approach [23]. In practice, an assistant model consists of a custom (more powerful) LLM in the sense that it is able to maintain context across multiple interactions within the same conversation.

The following prompt was used to retrieve the set of medicines corresponding to each description in the public procurement (in Portuguese):

> forneça uma tabela indexada contendo todos as substâncias, apresentações e códigos de barras de todas as linhas do arquivo cujo conteúdo inicial seja semelhante ao medicamento: "***desc_med***". As palavras na descrição do medicamento e nas linhas do arquivo podem estar embaralhadas, sem o espaçamento adequado, abreviadas ou conterem erros gramaticais. Desconsidere letras maiúsculas e minúsculas.

In the prompt, the term *desc_med* consists of the description of the medicine item contained in the public procurement.

5 Experiments and Analysis of the Results

5.1 Experimental Setup

For validating the proposed approaches, ten public procurements for medicines' purchasing from municipalities of Pernambuco state were analyzed. For short, we will refer to these public procurements as PN_1 to PN_{10}. These documents have 200, 105, 330, 356, 290, 238, 242, 169, 293, 70 medicines' descriptions, respectively.

The results obtained by the IR-Med were compared to an assistant model based on the GPT-4o with 128k context length. This assistant was developed to search in the same CMED table (i.e., the document containing the CMED items

[4] https://www.gov.br/anvisa/pt-br/centraisdeconteudo/publicacoes/medicamentos/publicacoes-sobre-medicamentos/vocabulario-controlado.pdf

was uploaded and processed by the GPT-4o model) of our proposed approach.

The performance of the methods will be assessed based on the following criteria: (i) accuracy on the information retrieval (i.e., percentage of elements in a public procurement whose all retrieved CMED rows truly correspond to the element description) and (ii) how reliable is the GPT-4o in terms of hallucination for this task?

Figure 3 depicts two examples of outputs scenarios. In Fig 3 a) the list of retrieved items also contains items that do not correspond to the provided description. This case is computed as an error of the method. The list of Fig 3 b) contains only items that match the description and it also a complete list (i.e., contains all possible items that match the description). Therefore, a correct retrieval respects the conditions described for Fig. 3 b). Nonetheless, the performance metrics (accuracy evaluation and hallucination) were manually computed by two practitioners.

5.2 Analysis of the Results

Table 2 depicts the results for each public procurement document and information retrieval method. The IR-Med approach successfully retrieved valid sets of CMED rows for over 80% of the public notices items' descriptions (for 7 out of 10 documents). Still, for the remaining documents, IR-Med yielded performances above 70%. On the other hand, the ChatGPT-4o assistant model retrieved an average accuracy of 30.9%. These values point out that the IR-Med method performed, in average, 2.85 times better than the LLM approach.

Table 2: Overall results of the IR-Med versus ChatGPT-4o assistant.

Public Notice	IR-MED	ChatGPT 4o assis.	GPT-4o Halluc. rate
PN_1	79.00 %	26.53 %	6.12 %
PN_2	86.92 %	36.92 %	6.92 %
PN_3	80.98 %	41.30 %	7.85 %
PN_4	86.52 %	39.19 %	6.63 %
PN_5	72.41 %	24.82 %	3.19 %
PN_6	81.51 %	22.08 %	6.06 %
PN_7	75.62 %	30.21 %	5.96 %
PN_8	84.62 %	38.75 %	8.13 %
PN_9	81.91 %	33.57 %	4.64 %
PN_{10}	84.29 %	15.94 %	7.25 %

The results of Table 2 present a lower bound in the IR-Med accuracy since in these values there are included cases where the description of the item is poorly organized (i.e., items in public procurement documents where it is not possible to find a corresponding list of medicines in the CMED list). This may occur due to a typo or the description of medicines associated to dosages (or forms) that are not available at the market, for example. These cases correspond to 6.00%, 1.54%, 3.61%, 1.69%, 5.17%, 4.20%, 4.13%, 3.55%, 4.10% and 2.86% of the items from public notices 1 to 10, respectively. These cases were also manually identified. So, in these situations the IR-Med approach can be used by both, audit practitioners and people in charge of creating the list of required medicines.

Table 2 also shows the percentage of items in each public procurement affected by the LLMs hallucination phenomenon [5, 10]. An hallucination case in this work takes place when the EAN number returned by the LLM is not a valid number (i.e., it is not present in the CMED list). The rates range from 3.19% to 8.13%. This is a particularly problematic aspect since that the lack of confidence in the results may affect the reputation of the solution. Another observed issue, however in smaller scale, is the occurrence of duplicate items in the retrieved set of rows. Therefore, a LLM based approach for the present problem clearly needs an extra processing phase in order to validate the retrieved information.

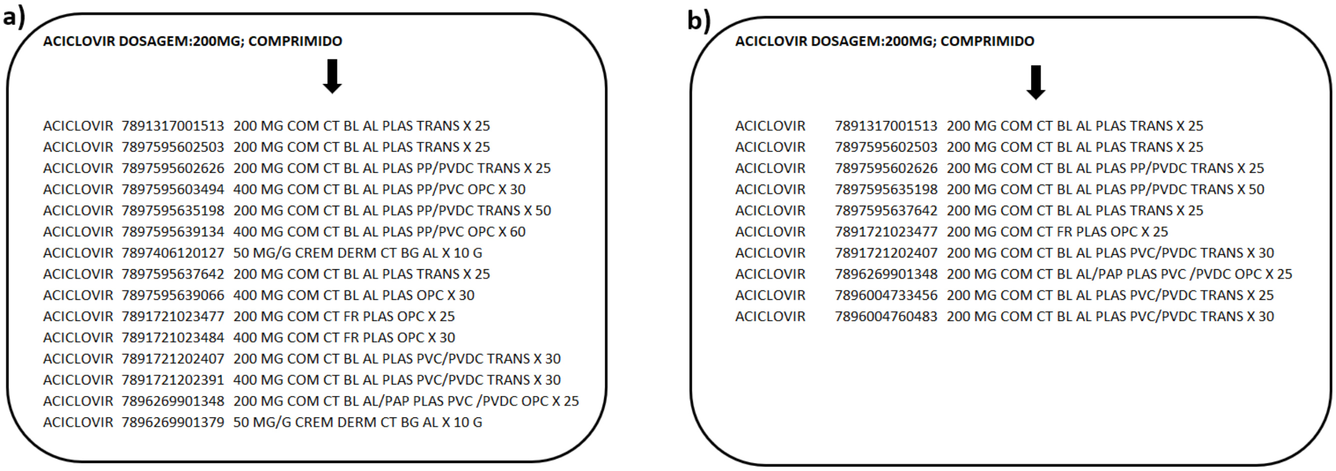
6 Threats to validity

Internal Validity: In order to offer reliable conclusions, our study investigated 10 different public procurements containing a total of 2293 medicine items. Furthermore, the moderately low standard deviation of the accuracies presented in Table 2 corroborates to the suitability of the depicted results.
Construct Validity: The main performance metric evaluated in this work was developed based on the expected behavior of an human audit practitioner. Therefore, the metric well represents the problem.
External Validity: this study was based on ten public procurement documents for medicine purchasing, as explained in Section 5. All of these documents are publicly available and cover a variety of different ways to describe medicines. Therefore, it is expected that the presented results generalize to other documents containing similar information.

7 Conclusion and Future Works

Auditing activities are inherently manual, i.e., non automated. However, parts of an auditing process can be automated in order to tackle the usual large amount of data. In this direction, examining public notices for medicines purchasing is a time consuming and error prone task. Since that for this problem a long list of medicines must be thoroughly examined, a human auditor can be affected by fatigue consequently lowering the quality of his/her work.

This work, in addition to a LLM based IR method, introduces an ad hoc method, referred to as IR-Med, capable of, given a non standardized description of a medicine, correctly identify a list of correspondent medicines in a public catalog of medicines available at the market and maintained by the national regulatory health agency. The results show that IR-Med is able to correctly identify, in most of cases, more than 80% of the medicines described in the public notice document. In addition, it also identifies cases were there is the need of further improvement in the medicine description. The work also showed that an LLM based approach might not be suitable since it consistently mix different medicines for a same item description. Still, it also suffers from complex problems such as hallucination and duplicated items retrieval.



a)

ACICLOVIR DOSAGEM:200MG; COMPRIMIDO

ACICLOVIR	7891317001513	200 MG COM CT BL AL PLAS TRANS X 25
ACICLOVIR	7897595602503	200 MG COM CT BL AL PLAS TRANS X 25
ACICLOVIR	7897595602626	200 MG COM CT BL AL PLAS PP/PVDC TRANS X 25
ACICLOVIR	7897595603494	400 MG COM CT BL AL PLAS PP/PVC OPC X 30
ACICLOVIR	7897595635198	200 MG COM CT BL AL PLAS PP/PVDC TRANS X 50
ACICLOVIR	7897595639134	400 MG COM CT BL AL PLAS PP/PVC OPC X 60
ACICLOVIR	7897406120127	50 MG/G CREM DERM CT BG AL X 10 G
ACICLOVIR	7897595637642	200 MG COM CT BL AL PLAS TRANS X 25
ACICLOVIR	7897595639066	400 MG COM CT BL AL PLAS OPC X 30
ACICLOVIR	7891721023477	200 MG COM CT FR PLAS OPC X 25
ACICLOVIR	7891721023484	400 MG COM CT FR PLAS OPC X 30
ACICLOVIR	7891721202407	200 MG COM CT BL AL PLAS PVC/PVDC TRANS X 30
ACICLOVIR	7891721202391	400 MG COM CT BL AL PLAS PVC/PVDC TRANS X 30
ACICLOVIR	7896269901348	200 MG COM CT BL AL/PAP PLAS PVC /PVDC OPC X 25
ACICLOVIR	7896269901379	50 MG/G CREM DERM CT BG AL X 10 G

b)

ACICLOVIR DOSAGEM:200MG; COMPRIMIDO

ACICLOVIR	7891317001513	200 MG COM CT BL AL PLAS TRANS X 25
ACICLOVIR	7897595602503	200 MG COM CT BL AL PLAS TRANS X 25
ACICLOVIR	7897595602626	200 MG COM CT BL AL PLAS PP/PVDC TRANS X 25
ACICLOVIR	7897595635198	200 MG COM CT BL AL PLAS PP/PVDC TRANS X 50
ACICLOVIR	7897595637642	200 MG COM CT BL AL PLAS TRANS X 25
ACICLOVIR	7891721023477	200 MG COM CT FR PLAS OPC X 25
ACICLOVIR	7891721202407	200 MG COM CT BL AL PLAS PVC/PVDC TRANS X 30
ACICLOVIR	7896269901348	200 MG COM CT BL AL/PAP PLAS PVC /PVDC OPC X 25
ACICLOVIR	7896004733456	200 MG COM CT BL AL PLAS PVC/PVDC TRANS X 25
ACICLOVIR	7896004760483	200 MG COM CT BL AL PLAS PVC/PVDC TRANS X 30

Figure 3: a) list of products containing correct and incorrect items w.r.t. the medicine description and b) list of products containing only correct items w.r.t. the medicine description.

A number of future works can be derived from the present study. For example: (i) investigating the public purchase of items other than medicines; (ii) improving the results achieved in the work and for a larger number of public notices; (iii) incorporate other phases of an auditing process such as the investigation of overpricing; (iv) combine different approaches such as LLMs and other methods in order to obtain more reliable and autonomous solutions; and (v) to use prompt engineering to avoid hallucination and improve the results quality.

Some limitations of this work are: (i) small number of experiments - given that the methods' performances are manually computed, carrying out a large number of experiments is an extremely time consuming task; and (ii) refinement of the use of LLMs - this work employed only one "fine-tuned" prompt for the used LLM. Many prompt engineering techniques can be investigated in the future further leveraging the LLMs potential for this problem.

Notice that the present work is highly related to some challenges present in the book “I GranDSI-BR – Grand Research Challenges in Information Systems in Brazil 2016-202” [6]. More specifically, this study is quite related to challenge 2 and chapters:

- *Information Systems and the Open World Challenges* - this study is partially connected to corruption issues which is an open challenge.
- *Information Systems based on (Linked) Open Data: From Openness to Innovation* - the developed approaches use transparency data to innovate an error prone process. In addition, it can be used not only by specialized people, its adoption by ordinary people is also important.
- *Transparency in Information Systems* - the developed approaches are entirely based on transparency data.

References

[1] Nora Abdelmageed and Sirko Schindler. 2021. JenTab Meets SemTab 2021's New Challenges.. In *SemTab@ ISWC*. 42–53.

[2] Abdulwahid Ahmad Hashed Abdullah and Faozi A. Almaqtari. 2024. The impact of artificial intelligence and Industry 4.0 on transforming accounting and auditing practices. *Journal of Open Innovation: Technology, Market, and Complexity* 10, 1 (2024), 100218. https://doi.org/10.1016/j.joitmc.2024.100218

[3] Ahmad Alobaid and Oscar Corcho. 2022. Balancing coverage and specificity for semantic labelling of subject columns. *Knowledge-Based Systems* 240 (2022), 108092.

[4] Ron Baker. 2019. Special issue: Government accounting, auditing and accountability: A Canadian perspective. *Canadian Journal of Administrative Sciences / Revue Canadienne des Sciences de l'Administration* 36, 2 (2019), 288–289.

[5] Dipto Barman, Ziyi Guo, and Owen Conlan. 2024. The Dark Side of Language Models: Exploring the Potential of LLMs in Multimedia Disinformation Generation and Dissemination. *Machine Learning with Applications* 16 (2024), 100545.

[6] Clodis Boscarioli, Renata Araujo, and Rita Suzana. 2017. *I GranDSI-BR Grand Research Challenges in Information Systems in Brazil 2016-2026 Organized by.*

[7] Michele A Brandão, Arthur PG Reis, Bárbara MA Mendes, Clara A Bacha de Almeida, Gabriel P Oliveira, Henrique Hott, Larissa D Gomide, Lucas L Costa, Mariana O Silva, Anisio Lacerda, et al. 2023. PLUS: A Semi-automated Pipeline for Fraud Detection in Public Bids. *Digital Government: Research and Practice* (2023).

[8] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2024. A Survey on Evaluation of Large Language Models. 15, 3, Article 39 (2024), 45 pages.

[9] Wenhu Chen, Ming-Wei Chang, Eva Schlinger, William Yang Wang, and William W. Cohen. 2021. Open Question Answering over Tables and Text. In *International Conference on Learning Representations*. https://openreview.net/forum?id=MmCRswl1UYl

[10] Adrian de Wynter, Xun Wang, Alex Sokolov, Qilong Gu, and Si-Qing Chen. 2023. An evaluation on large language model outputs: Discourse and memorization. *Natural Language Processing Journal* 4 (2023), 100024.

[11] Yves Emmanuel, Filipe Silva, George Cabral, and George Valença. 2023. Inovação na Contabilidade Pública - uma Solução que Analisa Atrasos de Pagamentos em Municípios Pernambucanos. 123–125. https://doi.org/10.5753/sbsi_estendido.2023.229382

[12] Hanchi Gu, Marco Schreyer, Kevin Moffitt, and Miklos Vasarhelyi. 2024. Artificial intelligence co-piloted auditing. *International Journal of Accounting Information Systems* 54 (2024), 100698.

https://doi.org/10.1016/j.accinf.2024.100698

[13] Jonathan Herzig, Thomas Müller, Syrine Krichene, and Julian Martin Eisenschlos. 2021. Open Domain Question Answering over Tables via Dense Retrieval. arXiv:2103.12011 [cs.CL] https://arxiv.org/abs/2103.12011

[14] Siqing Huo, Negar Arabzadeh, and Charles Clarke. 2023. Retrieving Supporting Evidence for Generative Question Answering. In *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*. Association for Computing Machinery, 11–20.

[15] Ensan F. Jafarzadeh, P. 2024. An evidence-based approach for open-domain question answering. *Knowledge and Information Systems* (2024). https://doi.org/10.1007/s10115-024-02269-2

[16] Emilia Kacprzak, José M Giménez-García, Alessandro Piscopo, Laura Koesten, Luis-Daniel Ibáñez, Jeni Tennison, and Elena Simperl. 2018. Making sense of numerical data-semantic labelling of web tables. In *Knowledge Engineering and Knowledge Management: 21st International Conference, EKAW 2018, Nancy, France, November 12-16, 2018, Proceedings 21*. Springer, 163–178.

[17] Xinyi Liang, Rui Hu, Yu Liu, and Konglin Zhu. 2024. Open-Domain Question Answering over Tables with Large Language Models. In *Advanced Intelligent Computing Technology and Applications*. 347–358.

[18] Jixiong Liu, Yoan Chabot, Raphaël Troncy, Viet-Phi Huynh, Thomas Labbé, and Pierre Monnin. 2023. From tabular data to knowledge graphs: A survey of semantic table interpretation tasks and methods. *Journal of Web Semantics* 76 (2023), 100761.

[19] Jixiong Liu, Viet-Phi Huynh, Yoan Chabot, and Raphaël Troncy. 2022. Radar station: Using kg embeddings for semantic table interpretation and entity disambiguation. In *International Semantic Web Conference*. Springer, 498–515.

[20] Sebastian Neumaier, Jürgen Umbrich, Josiane Xavier Parreira, and Axel Polleres. 2016. Multi-level semantic labelling of numerical values. In *The Semantic Web–ISWC 2016: 15th International Semantic Web Conference, Kobe, Japan, October 17–21, 2016, Proceedings, Part I 15*. Springer, 428–445.

[21] OpenAI. 2023. ChatGPT: Conversational Language Model. https://www.openai.com/chatgpt. Accessed: 2024-11-17.

[22] Arthur Silva, Vicente Sampaio, Adriano Lima, George Cabral, and George Valença. 2024. Ferramenta para Auxílio à Auditoria de Editais Municipais para Compra de Medicamentos. In *Anais Estendidos do XX Simpósio Brasileiro de Sistemas de Informação* (Juiz de Fora/MG). SBC, 265–268.

[23] Levy Silva and Luciano Barbosa. 2024. Improving dense retrieval models with LLM augmented data for dataset search. *Knowledge-Based Systems* 294 (2024), 111740.

[24] Sebastian Stephan, Johannes Lahann, and Peter Fettke. 2021. A Case Study on the Application of Process Mining in Combination with Journal Entry Tests for Financial Auditing. In *Hawaii International Conference on System Sciences*. https://doi.org/10.24251/HICSS.2021.694

[25] Svitlana Vakulenko and Vadim Savenkov. 2017. TableQA: Question Answering on Tabular Data. arXiv:1705.06504 [cs.IR] https://arxiv.org/abs/1705.06504

[26] Rafael B Velasco, Igor Carpanese, Ruben Interian, Octavio CG Paulo Neto, and Celso C Ribeiro. 2021. A decision support system for fraud detection in public procurement. *International Transactions in Operational Research* 28, 1 (2021), 27–47.

[27] Junjie Wang, Yuchao Huang, Chunyang Chen, Zhe Liu, Song Wang, and Qing Wang. 2024. Software Testing With Large Language Models: Survey, Landscape, and Vision. *IEEE Transactions on Software Engineering* 50, 4 (April 2024), 911–936.

[28] Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Jirong Wen. 2024. A survey on large language model based autonomous agents. *Frontiers of Computer Science* 18, 6 (2024). https://doi.org/10.1007/s11704-024-40231-1

[29] Yue Wang, Hung Le, Akhilesh Gotmare, Nghi Bui, Junnan Li, and Steven Hoi. 2023. CodeT5+: Open Code Large Language Models for Code Understanding and Generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). 1069–1088.

[30] Shuo Zhang and Krisztian Balog. 2020. Web table extraction, retrieval, and augmentation: A survey. *ACM Transactions on Intelligent Systems and Technology (TIST)* 11, 2 (2020), 1–35.

[31] Yiwei Zhou, Siffi Singh, and Christos Christodoulopoulos. 2021. Tabular data concept type detection using star-transformers. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 3677–3681.