Enhancing Dirty Data Detection via Large Language Models: Comparative Insights from ChatGPT, Gemini and Copilot

Damla Özyürek[a,b] and Aslı Suner[b, 1]

[a]Bandırma Onyedi Eylül University, Faculty of Medicine, Department of Biostatistics and Medical Informatics, Balıkesir, Turkey
[b]Ege University, Faculty of Medicine, Department of Biostatistics and Medical Informatics, İzmir, Turkey

DOI: 10.25929/ews67m03

ABSTRACT

Background: Dirty data, such as inconsistent categorical encodings, implausible values, duplicated records, and date/format errors, can compromise statistical inference and downstream decision-making. To support scalable data quality screening, LLMs are increasingly considered for data preparation tasks.

Methods: A tabular main dataset containing categorical and numerical variables was generated via stepwise prompting with ChatGPT and subsequently reviewed by domain experts. Seven representative data quality issues were injected (e.g., heterogeneous gender string variants, extreme BMI values due to decimal misplacement, gender–surgery incompatibilities, duplicates, admission date inconsistencies, character substitution [O↔0], and chronological/format errors). Using standardized prompts, the same datasets were uploaded to ChatGPT (v5.0), Gemini, and Copilot; queries were repeated across multiple attempts, and results were compared across models.

Results: Performance varied by error type. ChatGPT and Gemini struggled to consolidate semantically equivalent but differently written gender labels and produced incorrect frequency estimates, whereas Copilot identified more variants but still miscalculated proportions. All three models flagged extreme BMI outliers; Copilot, however, showed calculation errors in descriptive summaries. In detecting clinical plausibility constraints (e.g., gender–surgery mismatches) and structurally empty fields, models commonly misclassified empty cells as missing and, in some scenarios, generated hallucinated outputs. For duplicated records, ChatGPT and Gemini reliably detected duplicates but missed gender–surgery inconsistencies; Copilot's duplicate detection improved across attempts, and it more consistently captured gender-related mismatches. ChatGPT best identified admission date inconsistencies, while Gemini and Copilot failed to detect these errors. For character substitution, Gemini and Copilot detected the issue while ChatGPT did not. For chronological and date format errors, Copilot outperformed ChatGPT and Gemini.

Conclusions: LLMs can assist with rapid screening for certain data quality issues (notably outlier detection and some format-level anomalies), but they demonstrate inconsistent reliability for semantic normalization, rule-based clinical plausibility

[1] Corresponding author: Aslı SUNER, asli.suner@ege.edu.tr.

checks, and the interpretation of missingness. Human-in-the-loop verification and complementary deterministic validation rules remain essential when LLMs are used in data cleaning workflows.

KEYWORDS

Data quality, dirty data, data cleaning, large language models, ChatGPT, Gemini, Copilot, digital health

1. Introduction

High-quality clinical and administrative data are a prerequisite for reliable analytics and decision support; however, real-world datasets frequently contain missingness, inconsistencies, and format-level anomalies that are labor-intensive to detect and correct. Building on core data quality dimensions such as integrity, consistency, completeness, and accuracy, this study examines whether widely used large language models (LLMs) can serve as practical assistants for identifying common "dirty data" patterns in structured tabular datasets. Data integrity is defined as the completeness of the variables required for analysis and the conformity of the data to predefined formats. Data consistency refers to the absence of logically impossible values in the dataset and the coherence of data derived from related variables. Data integrity refers to the completeness of variables required for analysis and the conformity of data to predefined formats. Data consistency is defined as the absence of logically impossible values and the coherence of data derived from related variables. Data completeness and data accuracy involve identifying potential structural errors in the dataset through statistical analyses and ensuring the correctness of the data [1].

Dirty data are datasets that contain missing values, erroneous data entries, and inconsistencies [2]. Missing data, which are commonly encountered and negatively affect data quality, reduce statistical power and may lead to biased analyses [3]. Duplicate data increases frequency counts and introduces bias into analytical results. In medical datasets, duplicate records are frequently observed due to multiple patient encounters [4]. The presence of records in incompatible formats is common in clinical data and adversely affects data quality [5]. Unexpected outliers in variables may result from data entry errors, negatively impact data quality, and lead to misinterpretation of analytical results [6]. Recording data in different formats creates challenges during the analysis process; therefore, data are expected to conform to standardized formats during the data cleaning phase [7].

Data cleaning ensures that dirty data conform to data quality components and are suitable for statistical analyses. It improves the accuracy of analytical results, reduces bias, and enhances the reliability and reproducibility of findings [2]. The integration of data from multiple sources, along with the presence of missing values and inconsistent records, makes the data cleaning process particularly challenging [8]. Data cleaning is a time-consuming process that requires specialized expertise [9].

LLMs are actively involved in data analysis processes. Studies in the literature have examined their role as assistants in method selection and data analysis. Despite their advantages, they have limitations in performing advanced hypothesis testing. Expert oversight is therefore considered a requirement for statistical analyses conducted using LLMs [10]. They can save researchers considerable time in data cleaning and analysis tasks [11]. However, their capacity for data cleaning in text-based datasets remains limited [12]. In this study, we aim to evaluate the effectiveness of three different LLMs (ChatGPT, Gemini, and Copilot) in identifying errors within a dataset.

2. Methods

To comparatively evaluate the dirty data detection capabilities of three widely used LLMs, we designed a controlled, prompt-driven benchmarking workflow on a synthetic but clinically plausible tabular dataset. The main dataset was created iteratively with ChatGPT through a series of stepwise prompts designed in line with predefined variable types and coding schemes. Both the prompts and the generated outputs were reviewed by two independent domain experts. In addition, a third expert independently assessed the final dataset for overall correctness and identified potential data quality issues. All prompts were developed through consensus between two experts. For each scenario, the same modified dataset and a set of standardized prompts were used across ChatGPT 5.0, Gemini 2.5, and Copilot. Each model was run three times under identical conditions to assess consistency, sensitivity to different types of errors, and reproducibility during the study period (13–26 August 2026).

2.1 Main Dataset

The primary dataset, consisting of both categorical and continuous variables, was generated using ChatGPT through step-by-step prompts. The types of variables included in the dataset are presented in Table 1. The prompts and the generated data were reviewed by the experts involved in the study, and the accuracy of the results was independently evaluated by an external expert.

Categorical Variables	**Numerical Variables**
Gender (female, male)	Fasting Blood Glucose (mg/dL)
Surgical History (yes, no)	Age (year)
Surgery Type (cataract surgery, cholecystectomy, appendectomy, hernia repair, tonsillectomy, Cesarean section)	Height (cm)
	Body Weight (kg)
	Body Mass Index (kg/m^2)
	Pre-course quality of life scale score (1–10 points)
	Post-course quality of life scale score (1–10 points)
	Patient ID (6-digit ID number)

Table 1: Information for Variables

2.2 Prompts of LLMs and Data Quality Issues

For each application presented in Table 2, data quality issues consistent with the primary dataset were intentionally introduced. Using appropriate prompts, the dataset was uploaded to ChatGPT, Gemini, and Copilot, and each model was asked to identify the specified data quality issue. Within the primary dataset, inconsistencies between the variables gender and surgery type were present, including surgery diagnoses incompatible with patient sex. For each data quality query, the dataset was modified according to the data alterations specified in the table. The first data quality issue involved the presence of multiple data entries representing the category “female.” The second issue concerned the presence of outliers in the Body Mass Index variable due to data entry errors. The third issue consisted of surgery diagnoses incompatible with patient sex; specifically, Cesarean section procedures, which are not possible for male patients, were deliberately defined by the researcher. Duplicate records constituted the fourth data quality issue. The fifth issue involved the incorrect categorization of admission dates. In the sixth data quality issue, a character-level error was introduced by using the letter “O” instead of the numeral “0” (zero). The seventh data quality issue involved incorrect dates and inconsistencies in date formatting.

Data Quality Issues	**Prompts for LLMs**	**Data Quality Issues**
1	What is the number and percentage of females in this data set?	“Female”, “FEMALE”, “Kadın”, “KADIN”, “kadin” type inputs Gender-related data entry errors
2	Can you calculate descriptive statistics for “Body Mass Index” in this dataset?	Data entry error due to a misplaced decimal point Extreme values for BMI (269.27, 271.81) Gender-related data entry errors
3	Does this data set contain any incorrect or inappropriate data entries?	In two cases, a surgical diagnosis was recorded despite no surgery having been performed. Gender-related data entry errors (C/S for man, etc.)
4	Does this data set contain any incorrect or inappropriate data entries?	Duplicated data Gender-related data entry errors (C/S for man, etc.)

Data Quality Issues	Prompts for LLMs	Data Quality Issues
5	Does this data set contain any incorrect or inappropriate data entries?	Certain admission dates fall outside the past year; these records were still marked as "yes" (indicating admission within the last year). Gender-related data entry errors (C/S for man, etc.)
6	Does this data set contain any incorrect or inappropriate data entries?	Character substitution error (O ↔ 0) Gender-related data entry errors (C/S for man, etc.) Gender-related data entry errors (C/S for man, etc.)
7	Does this data set contain any incorrect or inappropriate data entries?	Chronological error (year 2034) Format error (day and month place exchange) Gender-related data entry errors (C/S for man, etc.)

Table 2: Data Quality Issues and Prompts for LLMs

2. Results

Table 3 presents the performance of the three LLMs in identifying different gender inputs. The models were tasked with detecting the frequency of the female category represented in the dataset using different lexical and formatting variations that convey the same meaning. It was observed that the ChatGPT and Gemini models failed to correctly identify formatting variations such as "Female," "FEMALE," or "Kadın," resulting in errors in frequency calculations. Copilot demonstrated comparatively better performance by recognizing multiple formatting variants; however, it still produced inaccuracies in frequency estimation.

Prompt	Dirty Data Problem	Expected Result	Number of Attempts	ChatGPT 5.0	Gemini	Copilot
What is the number and percentage of females in this data set?	"Female", "FEMALE", "Kadın", "KADIN", "kadin" type inputs Gender-related data entry errors (gender=male, surgery=Cesarean section)	50.20% female	1.	48.15% female	48.15% female	49.60% female
			2.	48.15% female	48.15% female	48.00% female
			3.	48.15% female	48.15% female	54.33% female
			Evaluation	**Failed to detect various forms of input** **Wrong calculation**		**Detection of various forms** **Wrong calculation**

Table 3: Performance of LLMs in Detecting and Quantifying Gender-Related Data Entry Variations

Table 4 presents the evaluation of extreme body mass index (BMI) values caused by misplaced decimal separators during data entry by the LLMs. ChatGPT and Gemini successfully identified the outliers across all three trials and accurately computed the descriptive statistics. In contrast, while Copilot issued warnings regarding the presence of outliers, it produced errors in the corresponding calculations.

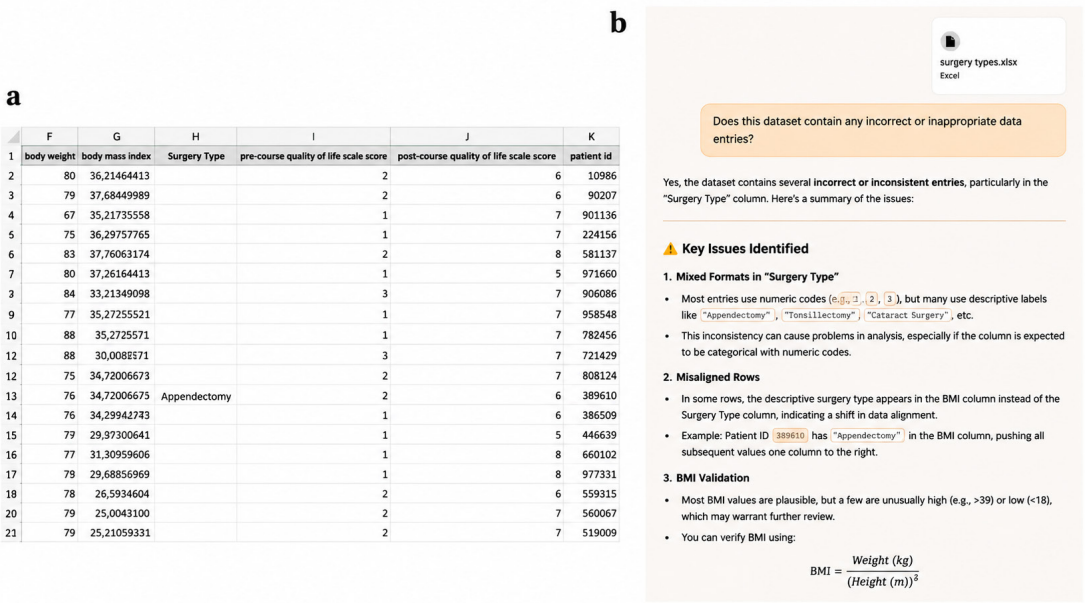
Prompt	Dirty Data Problem	Expected Result	Number of Attempts	ChatGPT 5.0	Gemini	Copilot
Can you calculate descriptive statistics for "Body Mass Index" in this dataset?	Data entry error due to a misplaced decimal point Extreme values for BMI (269.27, 271.81) Gender-related data entry errors (gender=male, surgery=Cesarean section)	Detection of extreme value	1.	Count: 243 Mean: 29.41 SD: 22.12 Variance: 489.35 Min: 21.80 25th percentile (Q1): 26.04 Median (Q2): 27.28 75th percentile (Q3): 28.98 Maximum: 271.81	Count: 243 Mean: 29.41 SD: 22.12 Minimum: 21.80 25th percentile (Q1): 26.04 Median (50th percentile):27.28 75th percentile (Q3): 28.98 Maximum:271.81	Count: 300 entries Mean: 27.38 Standard deviation: 11.26 Minimum: 21.80 Maximum: 271.81 Median: 26.67 25th percentile (Q1): 25.10 75th percentile (Q3): 28.69
			2.			
			3.			
			Evaluation	**Alert for extreme value**		**Alert for extreme value** **Wrong calculation**

Table 4: Performance of LLMs in Detecting and Analyzing Extreme BMI Values

Table 5 presents a comparison of model performance in detecting gender-related data entry errors and missing values. ChatGPT and Gemini failed to identify incorrect gender–surgery pairings and incorrectly classified cells that were intentionally left blank as missing data. Gemini reported no errors in the dataset during the first and last trials, whereas it incorrectly issued a missing data warning in the second trial. As illustrated in Figure 1, Copilot generated hallucinated outputs and misclassified cells that were intended to be empty as missing values. Although the dataset was correctly incorporated, the Copilot system incorrectly identified misaligned rows and referenced an ID that was not present in the data. This constitutes an instance of hallucination in LLMs.

Prompt	Dirty Data Problem	Expected Result	Number of Attempts	ChatGPT 5.0	Gemini	Copilot
Does this data set contain any incorrect or inappropriate data entries?	In two cases, a surgical diagnosis was recorded despite no surgery having been performed. (surgical history=no, surgery type=appendectomy) Gender-related data entry errors (gender=male, surgery=Cesarean section)	Inappropriate data entry detection of surgical type Detection of gender-related data entry errors	1.	Missing values: The Surgery Type column has 179 missing entries.	No noticeable anomalies	Both numeric codes and descriptive text entries in the "Surgical History" column Missing data Extreme BMI values
			2.		A few entries have a blank in "Surgery Type"	
			3.		No noticeable anomalies	
			Evaluation	**Erroneously flagged structurally empty fields as missing values** **Failed to detect gender-related data entry errors**	**Data error detection failed** **Erroneously flagged structurally empty fields as missing values** **Failed to detect gender-related data entry errors**	**Hallucination** **Erroneously flagged structurally empty fields as missing values** **Failed to detect gender-related data entry errors**

Table 5: Performance of LLMs in Detecting Inappropriate and Inconsistent Clinical Data Entries



	F	G	H	I	J	K
1	body weight	body mass index	Surgery Type	pre-course quality of life scale score	post-course quality of life scale score	patient id
2	80	36,21464413		2	6	10986
3	79	37,68449989		2	6	90207
4	67	35,21735558		1	7	901136
5	75	36,29757765		1	7	224156
6	83	37,76063174		2	8	581137
7	80	37,26164413		1	5	971660
8	84	33,21349098		3	7	906086
9	77	35,27255521		1	7	958548
10	88	35,2725571		1	7	782456
12	88	30,0088571		3	7	721429
12	75	34,72006673		2	7	808124
13	76	34,72006675	Appendectomy	2	6	389610
14	76	34,29942733		1	6	386509
15	79	29,97300641		1	5	446639
16	77	31,30959606		1	8	660102
17	79	29,68856969		1	8	977331
18	78	26,5934604		2	6	559315
20	79	25,0043100		2	7	560067
21	79	25,21059331		2	7	519009

Figure 1: Screenshot illustrating the dataset and an example of LLM hallucination output.

Table 6 presents the performance of the models in detecting duplicate records. Although ChatGPT and Gemini successfully identified repeated entries, they failed to detect gender-related inconsistencies in surgery types. Copilot did not identify duplicate records in the first trial but successfully detected them in subsequent trials. In contrast to the other models, Copilot consistently identified gender-related surgery type inconsistencies across all three trials. Across all three models, cells that were intentionally left blank were incorrectly interpreted as missing data.

<table>
<tr><th>Prompt</th><th>Dirty Data Problem</th><th>Expected Result</th><th>Number of Attempts</th><th>ChatGPT 5.0</th><th>Gemini</th><th>Copilot</th></tr>
<tr><td rowspan="4">Does this data set contain any incorrect or inappropriate data entries?</td><td rowspan="2">Gender-related data entry errors (gender=male, surgery=Cesarean section)

Duplicated data with two entries for each patient.</td><td rowspan="2">Detection of duplicated data

Detection of gender-related data entry errors</td><td>1.</td><td rowspan="2">A high number of duplicate entries

A large proportion of missing data in the Surgery Type variable entries</td><td rowspan="2">Duplicate rows

Missing values</td><td>Implausible or outlier values
Inconsistent surgery type entries
Missing or redundant data
Gender-specific mismatches</td></tr>
<tr><td>2.</td><td>Gender-specific surgery mismatches
Implausible age entries
BMI outliers
Missing or ambiguous surgery types
Duplicate or repetitive entries</td></tr>
<tr><td></td><td></td><td>3.</td><td></td><td></td><td>Gender-specific surgery mismatches
Implausible age entries
BMI outliers
Missing or ambiguous surgery types
Duplicate or repetitive entries</td></tr>
<tr><td></td><td></td><td>Evaluation</td><td colspan="2">Duplicated data detection

Failed to detect gender-related data entry errors

Erroneously flagged structurally empty fields as missing values</td><td>Duplicated data detection (attempt 2,3)

Detection of gender-related data entry errors

Erroneously flagged structurally empty fields as missing values

Failed to detect duplicated data (attempt 1)</td></tr>
</table>

Table 6: Performance of LLMs in Detecting Duplicate Records

Table 7 compares the models' abilities to detect data inconsistencies. Figure 2 presents representative screenshots of the date inconsistency detection process for ChatGPT and Gemini. Each model displays the data and reports its findings. The models were queried using a dataset in which the admission date fell outside the last year, yet the corresponding variable was incorrectly categorized as indicating an admission within the past year. ChatGPT successfully identified inconsistencies related to admission dates, whereas Gemini and Copilot failed to detect these errors. In addition, Gemini incorrectly classified cells that were intended to be empty as missing data, while Copilot emphasized that no errors were present in the relevant columns. Among the three models, ChatGPT demonstrated the best performance in detecting date-related errors. In this query, none of the three models was able to identify surgery diagnoses that were incompatible with patient gender.

Prompt	Dirty Data Problem	Expected Result	Number of Attempts	ChatGPT 5.0	Gemini	Copilot
Does this data set contain any incorrect or inappropriate data entries?	Certain admission dates fall outside the past year; these records were still marked as "yes" (indicating admission within the last year, for example). Reference date: 15 August 2025; admission date: 26 November 2023; admission within the past year: Yes (inconsistent). Gender-related data entry errors (gender=male, surgery=Cesarean section)	Data inconsistency detection	1. 2. 3.	Duplicate entries Missing data Admission date inconsistencies	Valid dates No duplicates Missing Values	Generally clean columns
			Evaluation	**Failed to detect gender-related data entry errors** **Detection data inconsistency**	**Erroneously flagged structurally empty fields as missing values** **Failed to detect gender-related data entry errors** **Failed to detect data inconsistency**	**Failed to detect gender-related data entry errors** **Failed to detect data inconsistency**

Table 7: Performance of LLMs in Detecting Date Inconsistencies

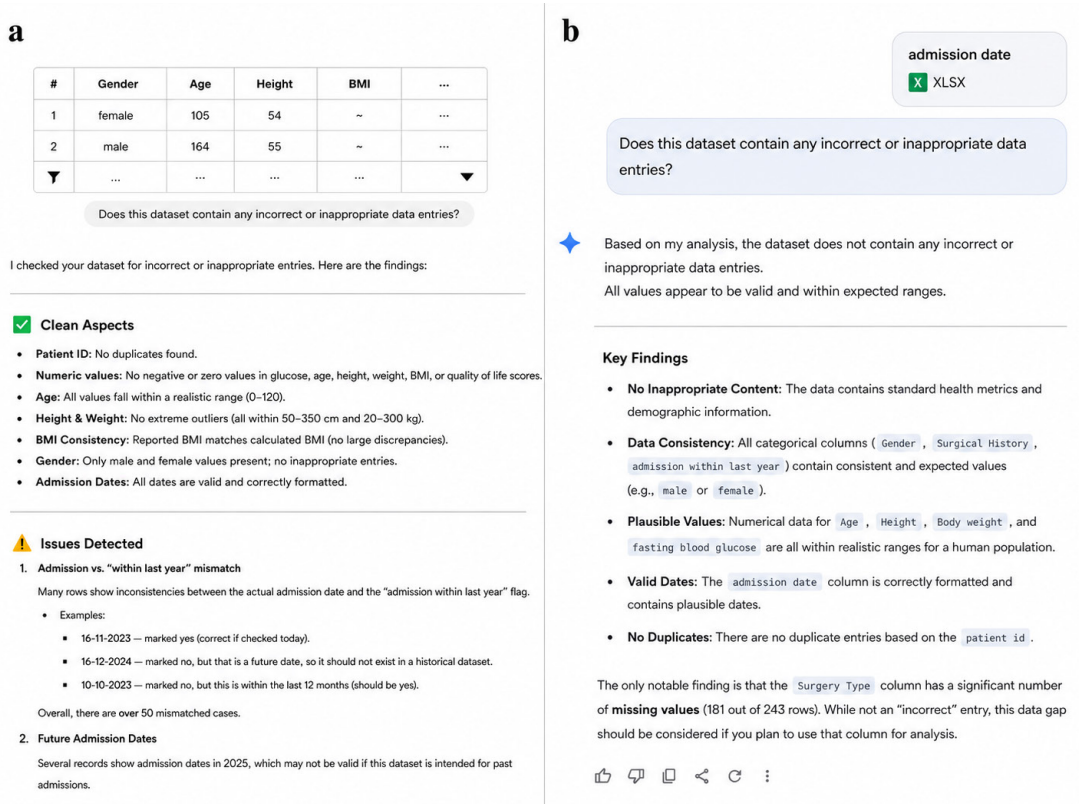


Figure 2: Screenshots for the data inconsistency detection of LLM (a. ChatGPT, b. Gemini).

Table 8 presents the performance of the LLMs in detecting character-level data entry errors. The dataset included instances in which the letter "O" was used in place of the numeral "0" (zero). While ChatGPT failed to detect this error, both Gemini and Copilot successfully identified it across all queries. ChatGPT incorrectly reported the presence of missing data and was unable to detect surgery diagnoses that were incompatible with patient gender. In addition to identifying character-level errors, Copilot demonstrated the strongest performance in detecting gender-incompatible surgery diagnoses.

Prompt	Dirty Data Problem	Expected Result	Number of Attempts	ChatGPT 5.0	Gemini	Copilot
Does this data set contain any incorrect or inappropriate data entries?	Character substitution error (O ↔ 0) Gender-related data entry errors (gender=male, surgery=Cesarean section)	Detection of character substitution error Detection of gender-related data entry errors (C/S for man, etc.)	1. 2. 3.	Missing values	Non-numeric character Incorrect data entries Incorrect data entries	Invalid patient IDs Biologically impossible surgical assignments Duplicate BMI values Outlier values
			Evaluation	**Failed to detect gender-related data entry errors** **Failed to detect character substitution error**	**Failed to detect gender-related data entry errors** **Detection of character substitution error**	**Detect gender-related data entry errors** **Detection of character substitution error**

Table 8: Performance of LLMs in Detecting Character-Level Data Entry Errors

Table 9 presents the models' performance in detecting date-related inconsistencies. The dataset included an impossible admission date (2034) and a date format mismatch, and the models were tasked with identifying these errors. ChatGPT and Gemini failed to detect either error across all three trials, whereas Copilot successfully identified both the invalid date and the date format inconsistency. In addition to failing to detect the errors, ChatGPT incorrectly issued missing data warnings. Although Copilot demonstrated the strongest performance in detecting this type of error, it also produced hallucinated outputs by generating results unrelated to the uploaded dataset.

Prompt	Dirty Data Problem	Expected Result	Number of Attempts	ChatGPT 5.0	Gemini	Copilot
Does this data set contain any incorrect or inappropriate data entries?	Chronological error (year 2034) Format error (day and month place exchange) Gender-related data entry errors (gender=male, surgery=Cesarean section)	Detect gender-related data entry errors Detect chronological error and format error	1. 2. 3.	Missing values No duplicates	Hallucination (different data set related findings)	Date format inconsistency Implausible admission data Gender-surgery Mismatches Outlier values
			Evaluation	**Failed to detect gender-related data entry errors** **Failed to detect chronological error and format error**	**Failed to detect gender-related data entry errors** **Failed to detect chronological error and format error**	**Detect gender-related data entry errors** **Detect chronological errors and format error**

Table 9: Performance of LLMs in Detecting Chronological and Date Format Errors

Discussion and Conclusion

Across seven common data-quality scenarios, encompassing both statistical and domain-specific logic errors, the three LLMs demonstrated complementary strengths alongside notable and consistent limitations. Overall, our findings suggest that LLM-based approaches can be useful for identifying more obvious anomalies, such as outliers and certain formatting inconsistencies. However, model performance was sensitive to how errors were represented (e.g., semantic versus syntactic forms) and was occasionally affected by miscalculations, misinterpretation of structurally empty fields, and sporadic hallucinations. These observations highlight the continued need for human oversight in clinical data-cleaning workflows.

LLMs may fail to evaluate semantically equivalent words within the same category due to differences in their orthographic representations [13]. In our study, ChatGPT and Gemini were unable to identify formatting variations such as "Female," "FEMALE," or "Kadın," whereas Copilot successfully detected these variants. This discrepancy can be attributed to differences in the models' sensitivity to surface-level formatting variations. LLMs are capable of detecting outliers in datasets [14]. Similarly, all three models in our study successfully identified outliers. Although Copilot detected the outliers, it made errors in the associated calculations. Although LLMs primarily operate by recognizing patterns within data, their capacity for medical reasoning remains limited [15]. In accordance with this limitation, none of the three models was able to detect inconsistencies between gender and surgery type during the initial queries. Previous studies have reported that LLMs can identify duplicate records during the data cleaning process [16], and all three models in our dataset demonstrated the same capability. LLMs are sensitive to missing data, and errors in missing data assessment have been previously reported [17]. Consistent with these findings, across different trials in our study, the models incorrectly classified cells that were intentionally left blank as missing data. For date-related inconsistencies, ChatGPT successfully detected errors associated with admission dates, whereas Gemini and Copilot failed to do so. Limitations in detecting data inconsistencies have been previously reported for LLMs [18], and our findings further support this evidence. Prior studies have demonstrated that LLMs exhibit limited performance in detecting character-level errors, such as character substitution, particularly in morphologically rich languages [19]. Our findings showed that Gemini and Copilot failed to detect character-level errors, whereas ChatGPT successfully identified such errors. This variation was attributed to differences in model sensitivity. Date-related errors and formatting inconsistencies are among the most commonly encountered issues in datasets [20]. Xu et al. [21] demonstrated that LLMs can effectively detect date, unit, numbering, and report template-related formatting errors in ophthalmic imaging reports. In our study, while ChatGPT and Gemini failed to identify date errors and formatting inconsistencies, Copilot successfully detected both. These findings suggest that the ability of these models to detect format- and date-related errors may depend on factors such as training data composition, contextual sensitivity, and error identification strategies.

In practice, LLMs should be positioned as assistive "first-pass" detectors rather than stand-alone validators, with results cross-checked through rule-based constraints (e.g., permissible category dictionaries, plausibility rules, and date-format validators) and expert review. While they offer rapid improvements in capability, they may still struggle with detection errors. Future studies should benchmark reproducibility across model versions and prompting strategies on real-world clinical datasets to clarify when LLMs add measurable value beyond conventional data quality tooling. These findings also point to several important directions for future research. First, the generation of benchmark datasets to assess the performance of different LLMs could help establish more standardized evaluation practices and provide a stronger foundation for their use in data cleaning processes. These datasets could be a standardized tool for researchers to test the capacity of new LLM tools in data cleaning tasks, and they serve as a tool to compare the capabilities of developing LLM tools. Second, attention should be given to the standardization of prompts used in LLM-assisted data cleaning, as well as to the validation of their effectiveness across different contexts. Standardized prompts can help produce more consistent outputs for every attempt at the same data cleaning task. Different data types are associated with distinct data quality issues. These data quality issues require appropriate handling methods. Interpreting standardized prompts on datasets with different structures and data types may help identify prompt workflows that are better suited to specific data-cleaning tasks. Researchers work with datasets of varying complexity, ranging from basic to highly structured forms. To use LLM tools effectively for data cleaning tasks, they require guidance and support. Educational resources that support researchers in designing specific prompts for a dataset may improve the effective and appropriate use of LLMs in practice. More advanced LLM tools may present accessibility challenges for researchers when attempting to integrate them into routine practice. The design of accessible and user-friendly LLMs capable of producing more consistent outputs across users would further strengthen their utility as assistive tools in data-cleaning workflows.

Acknowledgement

This study is derived from Damla Özyürek's ongoing master's thesis at Ege University Graduate School of Health Sciences, who is also a co-author of this work. This study was supported by TUBITAK (1002) – A Short Term Support Module (Project No: 324S310), titled "Designing an Artificial Intelligence Language Model–Supported Biostatistics Course to Promote Statistical Literacy and Reduce Statistics Anxiety," under the supervision of Assoc. Prof. Dr. Aslı Suner. Graduate student Damla Özyürek is supported as a scholarship fellow under this project and also receives support from TUBITAK in the 2210 – MSc/MA Scholarship Program.

References

[1] Pilowsky JK, Elliott R, Roche MA. Data cleaning for clinician researchers: Application and explanation of a data-quality framework. *Australian Critical Care*. 2024;37(5):827–833. https://doi.org/10.1016/j.aucc.2024.03.004.

[2] Guo M, Wang Y, Yang Q, et al. Normal Workflow and Key Strategies for Data Cleaning Toward Real-World Data: Viewpoint. *Interactive Journal of Medical Research*. 2023;12:e44310. https://doi.org/10.2196/44310.

[3] Sharifnia AM, Kpormegbey DE, Thapa DK, Cleary M. A Primer of Data Cleaning in Quantitative Research: Handling Missing Values and Outliers. *Journal of Advanced Nursing*. 2026;82(1):970–975. https://doi.org/10.1111/jan.16908.

[4] Goldberg EA, Ross CJ, Douglas VP, Ivanov A, Elze T, Miller JW, Lorch AC, IRIS Registry Analytic Center Consortium. Data Duplication and Errors in Large Medical Data Sets: A Case Study in the IRIS® Registry. *Ophthalmology Science*. 2025;6(1):100933. https://doi.org/10.1016/j.xops.2025.100933.

[5] Lewis AE, Weiskopf N, Abrams ZB, et al. (2023). Electronic health record data quality assessment and tools: a systematic review. *Journal of the American Medical Informatics* Association (JAMIA). 2023;30(10):1730–1740. https://doi.org/10.1093/jamia/ocad120.

[6] Syed R, Eden R, Makasi T, et al. Digital Health Data Quality Issues: Systematic Review. *Journal of Medical Internet Research*. 2023;25:e42615. https://doi.org/10.2196/42615.

[7] Penev YP, Buchanan TR, Ruppert MM, et al. Electronic Health Record Data Quality and Performance Assessments: Scoping Review. *JMIR Medical Informatics*. 2024;12:e58130. https://doi.org/10.2196/58130.

[8] Chicco D, Oneto L, Tavazzi E. Eleven quick tips for data cleaning and feature engineering. *PLoS Computational Biology*. 2022;18(12):e1010718. https://doi.org/10.1371/journal.pcbi.1010718.

[9] Martins P, Cardoso F, Váz P, Silva J, Abbasi M. Performance and Scalability of Data Cleaning and Preprocessing Tools: A Benchmark on Large Real-World Datasets. *Data*. 2025;10(5):68. https://doi.org/10.3390/data10050068.

[10] Huynh N, Lin B. LLMs Automatically Generate Code to Data Clean. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*. 2024.

[11] Fischer T, Biemann C. Exploring large language models for qualitative data analysis. In *Proceedings of the 4th international conference on natural language processing for digital humanities;* 2024 Nov. p. 423-437. https://doi.org/10.18653/v1/2024.nlp4dh-1.41.

[12] Meguellati E, Pratama N, Sadiq S, Demartini G. Are large language models good data preprocessors? In *Companion Proceedings of the ACM on Web Conference 2025*; 2025 May. p. 2129-2132. https://doi.org/10.48550/arXiv.2502.16790.

[13] Errica F, Sanvito D, Siracusano G, Bifulco R. What did I do wrong? Quantifying LLMs' sensitivity and consistency to prompt engineering. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies*. Volume 1: Long Papers; 2025 Apr. p. 1543–1558. https://doi.org/10.18653/v1/2025.naacl-long.73.

[14] Li A, Zhao Y, Qiu C, et al. Anomaly detection of tabular data using LLMs. arXiv [Preprint]. 2024. https://doi.org/10.48550/arXiv.2406.16308.

[15] Bedi S, Jiang Y, Chung P, Koyejo S, Shah N. Fidelity of medical reasoning in large language models. *JAMA Network Open*. 2025;8(8):e2526021–e2526021. https://doi.org/10.1001/jamanetworkopen.2025.26021.

[16] Zhang S, Huang Z, Wu E. Data cleaning using large language models. In *2025 IEEE 41st International Conference on Data Engineering Workshops (ICDEW)*. IEEE. 2025 May. p. 28–32. https://doi.org/10.1109/ICDEW67478.2025.00008.

[17] Kobayashi Y, Jeanselme V, Joshi S. (2025). Mind the data gap: Missingness Still Shapes Large Language Model Prognoses. *arXiv [Preprint]*. 2025. arXiv:2512.00479.

[18] Mahbub T, Dghaym D, Shankarnarayanan A, Syed T, Shapsough S, Zualkernan I. Can GPT-4 aid in detecting ambiguities, inconsistencies, and incompleteness in requirements analysis? A comprehensive case study. *IEEE Access*. 2024;12: 171972–171992. https://doi.org/10.1109/ACCESS.2024.3464242.

[19] Kwon SY, Bhatia G, Nagoud EMB, Abdul-Mageed M. ChatGPT for Arabic grammatical error correction. *arXiv [Preprint]*. 2023. https://doi.org/10.48550/arXiv.2308.04492.

[20] Hassenstein MJ, Vanella P. (2022). Data Quality—Concepts and Problems. *Encyclopedia*. 2022;2(1),498–510. https://doi.org/10.3390/encyclopedia2010032.

[21] Xu Y, Kang D, Shi D, Tham YC, Grzybowski A, Jin K. Leveraging ChatGPT for Report Error Audit: An Accuracy-Driven and Cost-Efficient Solution for Ophthalmic Imaging Reports. Ophthalmology and Therapy. 2025;1–14. https://doi.org/10.1007/s40123-025-01248-2.