Standardizing Genomic Reports: A Dataset, A Standardized Format, and A Prompt-Based Technique for Structured Data Extraction

Tamali Banerjee[1]*, Akshit Varmora[1]*, Jay Gorakhiya [1]*,
Sanand Sasidharan[2], Anuradha Kanamarlapudi[2], Pushpak Bhattacharyya[1]
[1]Computing for Indian Language Technology, IIT Bombay, India.
{tamali, akshitvarmora, jaygorakhiya, pb}@cse.iitb.ac.in
[2]GE Research, India.
{Sanand.Sasidharan, anuradha.kanamarlapudi}@gehealthcare.com

Abstract

Extracting information from genomic reports of cancer patients is crucial for both healthcare professionals and cancer research. While Large Language Models (LLMs) have shown promise in extracting information, their potential for handling genomic reports remains unexplored. These reports are complex, multi-page documents that feature a variety of visually rich, structured layouts and contain many domain-specific terms. Two primary challenges complicate the process: (i) extracting data from PDFs with intricate layouts and domain-specific terminology and (ii) dealing with variations in report layouts from different laboratories, making extraction layout-dependent and posing challenges for subsequent data processing.

To tackle these issues, we propose (a) GR-PROMPT, a prompt-based technique that uses a multimodal LLM to extract information from a genomic report, and (b) GR-FORMAT, a standardized format specifically designed to encapsulate all critical information within a genomic report in a structured manner. Together, these two convert a genomic report PDF of any layout into GR-FORMAT as a JSON file. This is the first approach to convert a genomic report PDF into a machine-readable, standardized format. To address the lack of available datasets for this task, we introduce GR-DATASET, a synthetic collection of 100 cancer genomic reports in PDF format. Each report PDF is accompanied by key-value information presented in a layout-specific format and structured key-value information in GR-FORMAT. This is the first dataset in this domain to promote further research for the task. We performed our experiment on this dataset. We publicly release[1] the code, the format, and data for further research.

1 Introduction

Genomic reports provide detailed insights into oncology patients' biomarkers, including specific genetic mutations and associated therapies. These reports primarily assist healthcare professionals in devising treatment strategies for cancer patients. The increasing workload on healthcare professionals heightens the risk of human error, further driving the demand for AI-driven assistance to help them make faster and more accurate treatment decisions. Additionally, the information gathered from these reports can contribute to future cancer research. This underscores the need for efficient and accurate systems to extract and interpret data from genomic reports.

The problem addressed in this study is the development of an automated system capable of extracting and interpreting key information from genomic reports.

Input: The PDF file of the Genomic Report for oncology patients.

Output: The JSON file containing information of the report in GR-FORMAT.

One of the primary challenges in processing genomic reports lies in their inherent complexity. While these reports contain structured data, complexity often arises from the distribution of information across multiple pages, where key data points may be interrupted by page breaks or dispersed throughout non-adjacent sections. This design is presumably intentional, as the layouts of genomic reports in certain laboratories prioritize critical information for human healthcare professionals. By placing the most pertinent data on the first page, the reports aim to facilitate quick access and reduce the time spent searching for essential information. However, this emphasis on accessibility leads to a more intricate PDF layout, complicating the data extraction process. However, for machine understanding, all pages hold equal visibility, and

*These authors contributed equally to this work and are jointly the first authors.

[1] https://github.com/akshit-1729/IE-GR

machines prefer data to be organized consistently for effective processing. Additionally, using visual elements and domain-specific entities (highly specialized contents that need a better contextual understanding of the domain) adds to this complexity. Another challenge is the diversity of report layouts used by different laboratories, as there is no standardized format for these reports. This variability complicates automated information extraction, as subsequent steps should not be biased toward specific laboratory formats.

To address these challenges, we propose GR-FORMAT, a standardized JSON format that can be used to convert any genomic report, making it layout-agnostic, regardless of the original layout. We developed this format by analyzing genomic reports with varying layouts. Additionally, we propose GR-PROMPT, a prompt-based technique to extract information from genomic report PDFs in GR-FORMAT. GR-PROMPT has a set of prompt which utilizes the GPT-4 (Achiam et al., 2023), a multi-modal closed-source LLM to do the task.

There is no publicly available dataset for this task, as publishing genomic reports contains privacy and consent issues. In response, we introduce the GR-DATASET, a synthetic dataset consisting of 100 cancer genomic reports, key-value information presented in a layout-specific format, as well as structured key-value information in GR-FORMAT.

This research presents three key contributions:

1. **GR-FORMAT:** A standardized JSON format specifically designed to encapsulate all critical information within a genomic report in a structured manner. This is the first such standardized format in this domain, ensuring consistency and eliminating layout variability across reports from different laboratories. By converting genomic reports into GR-FORMAT, subsequent processes like clinical trial matching or treatment recommendation modules can easily retrieve and process key data without concern for differing report layouts.

2. **GR-PROMPT:** A prompt-based approach leveraging multimodal LLM to extract and convert genomic reports, typically in PDF format, into a standardised JSON format. This is the first approach to convert a genomic report PDF into a machine-readable, standardized format. We achieved an overall accuracy of 73.19% using this technique.

3. **GR-DATASET:** First publicly available dataset of synthetic genomic reports, comprising 100 synthetic cancer genomic reports in PDF format. Each report is accompanied by key-value information presented in a layout-specific format, as well as structured key-value information in GR-FORMAT. This dataset is designed as a test set for evaluating information extraction methods from genomic reports, with the goal of advancing research and development in the field.

2 Related Work

A related work that processes genomic reports' data is GENETEX (Miller and Shalhout, 2021). It is a tool that converts semi-structured data to structured data using text-mining and regular expressions. However, to the best of our knowledge no work extracts key information from a genomic report and outputs it in a predefined format. However, they did not release the dataset publicly to further facilitate research in research on Genomic report information extraction.

Some works have been done to extract structured information from text using LLM (Dagdelen et al., 2024; Wu et al., 2024). (Dagdelen et al., 2024) highlights the ability of fine-tuned LLMs, like GPT-3 and Llama-2, to extract intricate relationships from scientific literature, enabling flexible output formats for database creation. (Wu et al., 2024) introduces a novel entity-centric approach called Structured Entity Extraction (SEE), which utilizes the Approximate Entity Set OverlaP (AESOP) metric for performance evaluation and demonstrates enhanced extraction efficiency through a multi-stage model, MuSEE.

Another direction of work that closely resembles our work is extracting information from PDFs that have complex layouts. Donut (Kim et al., 2022) is a model designed for document image understanding that leverages a unified architecture to process both text and layout information effectively, enabling it to handle diverse document types without the need for explicit layout annotations. DocOwl 1.5 (Ye et al., 2023) is a versatile tool that combines advanced parsing capabilities with intelligent chunking strategies, ensuring that hierarchical relationships within documents are preserved during extraction, thus enhancing the accuracy and usability of the extracted data. LayoutLMv3 (Huang et al., 2022) enhances document layout comprehension by integrating unified text and image masking techniques, allowing it to excel in tasks like form understanding and visual question answering. LayoutLLMs (Luo et al., 2024; Fujitake, 2024) build their systems by focusing on leveraging large language models for improved contextual understanding of document structures, making them adept at extracting relevant information from intricate layouts.

A recent work (Tam et al., 2024) emphasizes that while LLMs are powerful tools, their performance can be significantly affected by the constraints imposed on their output formats, necessitating careful consideration in practical applications.

3 Methodology

In this section, we describe GR-FORMAT (the standardised format), GR-PROMPT (the prompt-based approach), and GR-DATASET (the dataset) in detail.

3.1 The Specified Format: GR-FORMAT

The extracted output from the genomic report must utilize a layout-agnostic JSON format to facilitate downstream processing. For example, one layout presents associated therapies alongside biomarkers at the beginning of the report, while another layout separates biomarkers from their corresponding therapies and includes a mapping to illustrate their relationships. These variations likely aim to prioritize critical information for healthcare professionals, ensuring that the most important data is easily accessible on the first page, thereby reducing time spent searching. However, for machine understanding, all pages hold equal visibility, and machines prefer data to be organized consistently for effective processing. This approach ensures that biomarkers, mutations, and associated data are consistently organized, thereby preventing complications and minimizing the risk of overlooking critical information.

To address the variability in genomic report layouts, we developed a standardized format that effectively captures all relevant key information, irrespective of the original layout. For example, this format accommodates diverse layouts while maintaining a coherent structure for data extraction. This format was created by analyzing reports with diverse layouts and defining a common set of keys. It can be represented as either a JSON object or a hierarchical list. Our experiment uses the format as a hierarchical list for easy conversion, avoiding strict format like JSON as suggested by (Tam et al., 2024). However, we ultimately convert it to JSON to ensure the final output is machine-readable. On the other hand, the GR-DATASET stores corresponding key-value pairs of genomic reports in GR-FORMAT as JSON. Therefore, we evaluate the accuracy of the final JSON output by comparing it with the JSON of the dataset. The standardized structure consists of eight sections, each of which is designed to serve a specific purpose.

1. Patient Information: It captures essential details about the patient, including identifiers and contact information, which are crucial for linking the genomic data to the individual.

2. Diagnosis Information: It provides information related to the diagnosis process, including the diagnosing center, doctor, and laboratory methods used, ensuring transparency in how the diagnosis was reached.

3. Cancer Information: It lists the details such as the type, stage, and grade of cancer, facilitating an understanding of the cancer's characteristics and biological context.

4. Biomarkers: Biomarkers can be of different types, including gene mutations.

   (a) General Biomarkers: Includes metrics like Microsatellite Instability (MSI) and Tumor Mutational Burden (TMB), which are critical for assessing the tumor's biology and potential treatment responses.

   (b) Gene Mutations: For each mutation, relevant details such as gene name, mutation type, and pathogenicity are provided, offering insight into the genetic alterations present in the tumor.

   (c) Immunochemistry Biomarkers: Focuses on specific biomarkers relevant to immunotherapy, detailing expression levels and interpretations.

5. Therapeutic Information: It lists therapies that are approved by FDA (Food and Drug Administration) associated with the cancer type and related biomarkers.

   (a) FDA-Approved Therapies for Current Diagnosis: It lists therapies specifically

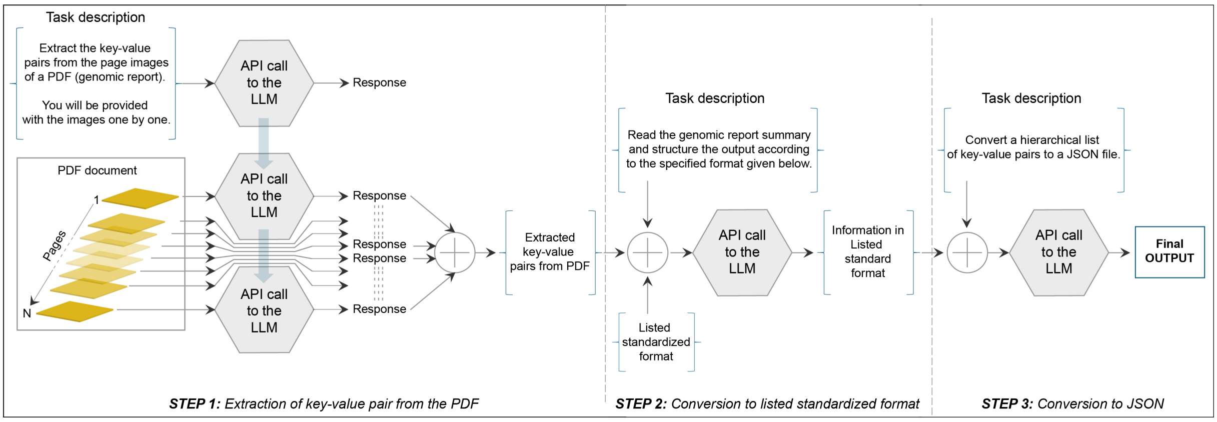

Task description
Extract the key-value pairs from the page images of a PDF (genomic report).
You will be provided with the images one by one.
API call to the LLM
Response
PDF document
1
Pages
N
API call to the LLM
Response
Response
Response
API call to the LLM
Response
Extracted key-value pairs from PDF
Task description
Read the genomic report summary and structure the output according to the specified format given below.
API call to the LLM
Information in Listed standard format
Listed standardized format
Task description
Convert a hierarchical list of key-value pairs to a JSON file.
API call to the LLM
Final OUTPUT
STEP 1: Extraction of key-value pair from the PDF
STEP 2: Conversion to listed standardized format
STEP 3: Conversion to JSON

Figure 1: Our prompt-based approach: GR-PROMPT. Thick arrows indicate conversation flow and the '+' sign represents concatenation.

approved for the diagnosed cancer, including associated biomarkers that may guide treatment decisions.

(b) FDA-Approved Therapies for Other Indications: It identifies therapies that might apply to other conditions, along with relevant biomarkers.

6. Clinical Trials: It shows ongoing clinical trials relevant to the patient, including trial titles, phases, and associated mutations, which can provide opportunities for advanced treatment options.

7. Variants of Unknown Significance: It catalogs variants that lack established significance, detailing their characteristics to help guide further investigation and clinical decisions.

8. Additional Indicators: It includes prognostic markers, other molecular indicators, and special notes regarding cancer progression or drug resistance, providing a comprehensive view of the patient's condition and potential treatment challenges.

3.2 The prompt-based approach: GR-PROMPT

Genomic reports are complex not only for their highly domain-specific contents but also for the use of visually-rich elements with color information. To include visual information with textual information, our approach uses a multimodal LLM to extract key-value pairs from genomic reports. Figure 1 shows our prompt-based approach for extracting information from a genomic report PDF into a standardised JSON format. In general, the approach includes the following steps with prompts.

1. Extraction of key-value pairs from the PDF: First, we extract relevant data points from the PDF report as key-value pairs.

   (a) Task Definition: This step defines the task for the LLM, which involves extracting key-value pairs from sequentially provided images of a genomic report PDF.

   (b) Sequential Input Processing: For each page, the LLM extracts key-value pairs while maintaining context across multiple pages as conversation history. The extracted data is then compiled into a unified summary.

2. Conversion to listed standardized format: In this step, we convert the extracted key-value pairs in our proposed format GR-FORMAT as a structured, hierarchical list. The inputs for this step include the unified key-value pairs from the entire PDF and the standardized format represented as a hierarchical list. Please note, we do not use JSON format here to avoid strict format like JSON as suggested by (Tam et al., 2024).

3. Conversion to JSON: Lastly, we convert the hierarchical list data into JSON for easy machine readability. The input for this step is the output from the previous stage, which consists of information formatted as a hierarchical list in GR-FORMAT.

We use multiple prompts to do the task instead of using a single prompt to convert it directly. It improves the output quality at the expense of cost and runtime.

3.3 The Synthetic Dataset: GR-DATASET

Medical datasets, especially cancer genomic reports data, are very scarce in number due to privacy issues and the sensitive information they contain. In this project, we address the unavailability of genomic reports by generating synthetic cancer genomic reports. The key challenges in creating synthetic data for the domain of genomic reports are the document formats' complexity and the document layouts' diversity. To overcome these challenges, we leverage publicly available data, specifically a comprehensive list of terminologies and options for each report element provided by the National Cancer Institute Thesaurus (NCIT) (nci). Each entry of the dataset consists of 3 parts. These are (i) genomic report in pdf format, (ii) key-value information presented in a layout-specific format as JSON file, and (iii) key-value information in GR-FORMAT as a JSON file. Synthetic generation of cancer genomic reports includes the following steps.

- Create one JSON file format for each layout
- Generate a JSON file with synthetic report data
- Create a visually rich layout
- Generate the final report by inserting the data from JSON file into the generated layout

We apply a rule-based method to map the genomic report information into GR-FORMAT and generate a corresponding JSON file, which generates the JSON file with synthetic data using dictionaries of the keywords (see section 4.2). The final document is created by populating the visually rich layout using the synthetically generated JSON file.

4 Experimental Setup

In this section, we provide the experimental details for applying GR-PROMPT to extract information from genomic reports in GR-FORMAT on the GR-DATASET. Additionally, we describe the process of generating synthetic cancer genomic reports and outline the evaluation strategy used to assess the quality of the output when applying GR-PROMPT to extract information from genomic reports in GR-FORMAT on the GR-DATASET.

4.1 Information Extraction from Genomic Reports

In our experiment, we utilized the GPT-4o model as our multimodal LLM, chosen for its advanced capabilities in understanding and processing diverse textual inputs.

To optimize the performance of GR-PROMPT, we configured two parameters. We set the maximum token limit to 4,000 to ensure that the model can process the entirety of each genomic report without exceeding the token constraints. This configuration allows the model to retain relevant context from the input data, which is critical for accurate key-value pair extraction.

Additionally, we adjusted the temperature setting to 0.5. This lower temperature value was selected to promote a more focused and deterministic response from the model, reducing variability and enhancing the precision of the outputs. By fine-tuning these parameters, we aimed to create a controlled environment that maximizes the reliability and relevance of the extracted information.

4.2 Synthetic Dataset Creation as Test-data

We tried three methods for generating visually rich report layouts and thus the final document: Microsoft Word, docx-mailmerge [2], Python Docx [3], and Python Reportlab [4]. For a detailed comparison of different methods, refer to Table 1. Comparative analysis showed that Python Reportlab outperforms the other methods in terms of customization, visual quality, and ease of use, making it the preferred approach for synthetic data generation in this project.

Medical dictionaries for genes, protein variants, cancer types, and therapies were also created to ensure accurate synthetic data generation (nci; fda). For number of entries in these dictionaries refer Table 2. We also use a list of therapies containing 1523 entries. We have created genomic reports of 2 different layouts. We have created 50 reports for each layout making a total of 100 visually rich genomic reports. Reports with Layout-1 contain 3 pages per document and reports with Layout-2 contain 5 pages per document. The first pages of both layouts are shown in Figure 2 and Figure 3, respectively.

4.3 Evaluation Strategy

To assess the effectiveness of our approach in extracting structured information from genomic reports, we developed a comprehensive evaluation framework. This framework compares the output

[2] https://pypi.org/project/docx-mailmerge2/

[3] https://python-docx.readthedocs.io/en/latest/

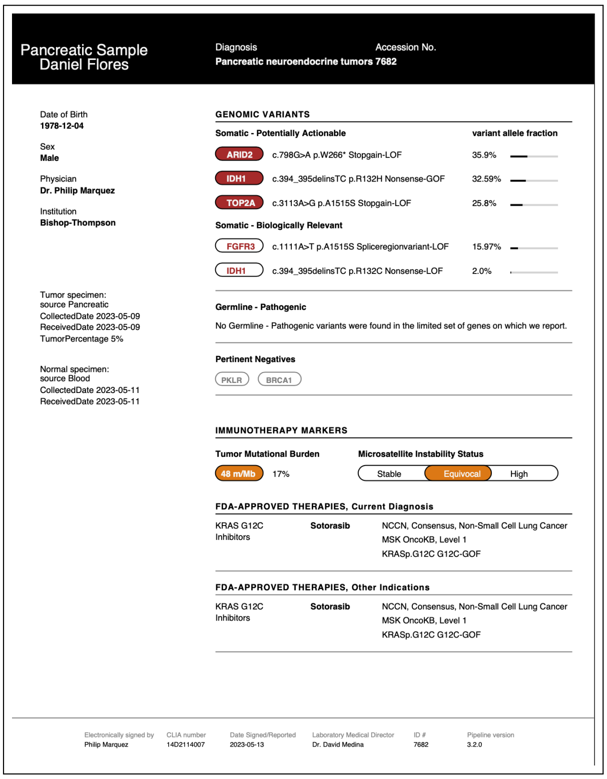
[4] https://docs.reportlab.com/

Method	Description
Microsoft Word, docx-mailmerge	Automates data insertion but lacks flexibility for dynamic fields.
Python Docx	Allows programmable layout creation but struggles with complex layouts.
Python Reportlab	Dynamically generates highly customizable and visually rich reports, including complex elements like graphs and tables.

Table 1: Methods for Generating Visually Rich Report Layouts

Dictionary	#of entries
Gene : Protien variant	650
Gene : Gene mutation	604
Cancer type : Diagnosis	39
Cancer type : Specimen type	37

Table 2: Medical terminologies dictionary statistics



Pancreatic Sample
Daniel Flores

Diagnosis
Pancreatic neuroendocrine tumors

Accession No.
7682

Date of Birth
1978-12-04

Sex
Male

Physician
Dr. Philip Marquez

Institution
Bishop-Thompson

Tumor specimen:
source Pancreatic
CollectedDate 2023-05-09
ReceivedDate 2023-05-09
TumorPercentage 5%

Normal specimen:
source Blood
CollectedDate 2023-05-11
ReceivedDate 2023-05-11

GENOMIC VARIANTS

Somatic - Potentially Actionable		variant allele fraction
ARID2	c.798G>A p.W266* Stopgain-LOF	35.9%
IDH1	c.394_395delinsTC p.R132H Nonsense-GOF	32.59%
TOP2A	c.3113A>G p.A1515S Stopgain-LOF	25.8%
Somatic - Biologically Relevant		
FGFR3	c.1111A>T p.A1515S Spliceregionvariant-LOF	15.97%
IDH1	c.394_395delinsTC p.R132C Nonsense-LOF	2.0%

Germline - Pathogenic

No Germline - Pathogenic variants were found in the limited set of genes on which we report.

Pertinent Negatives

PKLR BRCA1

IMMUNOTHERAPY MARKERS

Tumor Mutational Burden
48 m/Mb 17%

Microsatellite Instability Status
Stable [✓ Equivocal] High

FDA-APPROVED THERAPIES, Current Diagnosis

KRAS G12C Inhibitors	**Sotorasib**	NCCN, Consensus, Non-Small Cell Lung Cancer MSK OncoKB, Level 1 KRASp.G12C G12C-GOF

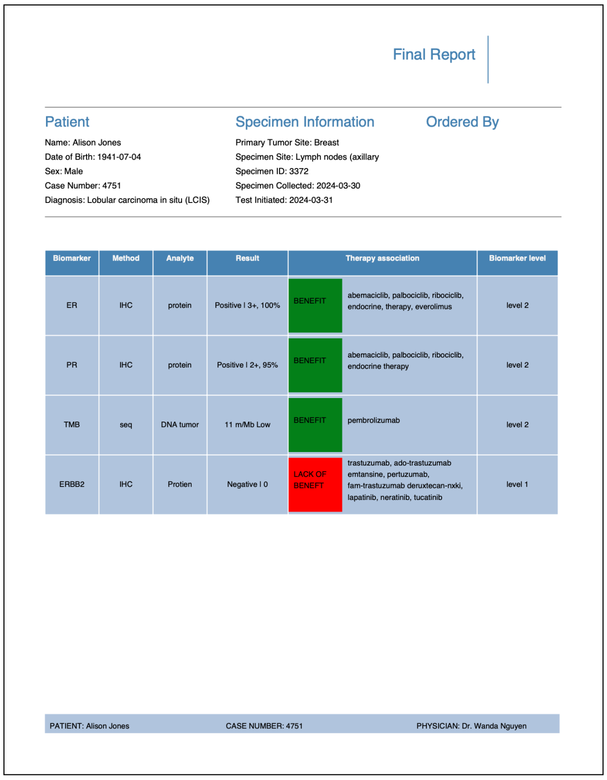
FDA-APPROVED THERAPIES, Other Indications

KRAS G12C Inhibitors	**Sotorasib**	NCCN, Consensus, Non-Small Cell Lung Cancer MSK OncoKB, Level 1 KRASp.G12C G12C-GOF

Electronically signed by	CLIA number	Date Signed/Reported	Laboratory Medical Director	ID #	Pipeline version
Philip Marquez	14D2114007	2023-05-13	Dr. David Medina	7682	3.2.0

Figure 2: Sample first page of a report with Layout-1

against the ground truth dataset using the comparison strategy mentioned in section 4.3.1. The ground truth, in this case, is the synthetic data we generated during the creation of our dataset described in section 4.2, which represents the ideal extraction of information from the genomic reports.



Final Report

Patient
Name: Alison Jones
Date of Birth: 1941-07-04
Sex: Male
Case Number: 4751
Diagnosis: Lobular carcinoma in situ (LCIS)

Specimen Information
Primary Tumor Site: Breast
Specimen Site: Lymph nodes (axillary
Specimen ID: 3372
Specimen Collected: 2024-03-30
Test Initiated: 2024-03-31

Ordered By

Biomarker	Method	Analyte	Result	Therapy association		Biomarker level
ER	IHC	protein	Positive I 3+, 100%	BENEFIT	abemaciclib, palbociclib, ribociclib, endocrine, therapy, everolimus	level 2
PR	IHC	protein	Positive I 2+, 95%	BENEFIT	abemaciclib, palbociclib, ribociclib, endocrine therapy	level 2
TMB	seq	DNA tumor	11 m/Mb Low	BENEFIT	pembrolizumab	level 2
ERBB2	IHC	Protien	Negative I 0	LACK OF BENEFT	trastuzumab, ado-trastuzumab emtansine, pertuzumab, fam-trastuzumab deruxtecan-nxki, lapatinib, neratinib, tucatinib	level 1

PATIENT: Alison Jones CASE NUMBER: 4751 PHYSICIAN: Dr. Wanda Nguyen

Figure 3: Sample first page of a report with Layout-2

4.3.1 Comparison Strategy

Our comparison strategy involves a detailed, hierarchical analysis of the JSON structures representing both the ground truth and the AI-generated output. The core of this strategy is implemented such that it performs a recursive comparison of these JSON objects. Here is an overview of the key components of our comparison strategy:

- **Recursive Comparison:** We traverse both JSON structures simultaneously, comparing each key-value pair at every level of the hierarchy.
- **Case-Insensitive Key Matching:** To account for minor variations in key naming, we perform case-insensitive comparisons of dictionary keys.

- **Value Normalization:** Before comparing values, we normalize them to account for different representations of the same data. This includes handling variations in data types (e.g., integers vs. strings), formatting (e.g., lists vs. comma-separated strings), synonymous key names and abbreviations.

- **Flexible List Comparison:** For certain fields like "Gene Mutations", we compare lists based on their content rather than their order, allowing for flexibility in the output structure.

5 Results

Table 3 presents the accuracy of the GR-PROMPT output in extracting information from genomic reports formatted in GR-FORMAT using the GR-DATASET. We display the accuracies for the two layouts separately.

	Layout-1	Layout-2
Accuracy	77.62%	68.76%

Table 3: Accuracy of GR-PROMPT approach on Layout-1 and Layout-2

5.1 Analysis

5.1.1 Quantitative Analysis

The overall accuracy of **73.19%** indicates that the model successfully extracted and correctly structured nearly three-fourth of the information from the genomic report. While this demonstrates the model's ability to capture a significant portion of the report's content, it also highlights areas for improvement.

5.1.2 Error Analysis

Examining the incorrect pairs reveals patterns in the types of errors the model makes:

- **Structural Mismatches:** Some fields expected to be lists were extracted as strings, or vice versa. For example, "Method of Analysis" was extracted as a string instead of a list.

- **Formatting Inconsistencies:** Minor differences in formatting, such as the presence or absence of percentage signs in "Variant Allele Fraction", led to mismatches.

- **Content Errors:** In some cases, the model extracted incorrect information. For example, in the "Gene Mutations" section, certain gene names and their associated data were either mismatched or missing. In Layout-2, where gene mutation data is spread across multiple tables, the model occasionally fails to accurately identify and consolidate the information from all the relevant tables.

- **Granularity Issues:** Some fields, like "Cancer Grade", were partially correct but lacked the full detail present in the ground truth.

5.1.3 Field-Specific Performance

Analyzing the performance across different sections of the report reveals varying levels of accuracy:

- **Patient and Diagnosis Information:** The model performed well in extracting basic patient details and diagnosis information, with only minor discrepancies (e.g., a one-year difference in patient age).

- **Cancer Information:** While the model captured the main cancer type correctly, it struggled with the granularity of information in fields like "Cancer Grade" and "Tumor Specimen Source".

- **Biomarkers:** The extraction of biomarker information showed mixed results. General biomarkers were mostly correct, but the "Gene Mutations" section had significant discrepancies.

- **Therapeutic Information:** The model generally captured the types of therapies correctly but sometimes missed associated biomarkers.

- **Clinical Trials:** Information about clinical trials was largely correct, with some minor formatting differences in medication lists.

6 Summary, Conclusion, and Future work

In this work, we introduced GR-PROMPT, a novel technique for extracting information from complex cancer genomic reports, and GR-FORMAT, a standardized data format, using a multimodal LLM to handle intricate layouts and domain-specific terminology. We also introduced GR-DATASET, the first synthetic dataset for this task, facilitating future research.

While our approach demonstrates a promising trajectory, future work will focus on expanding the dataset to improve the model's performance and adaptability across diverse healthcare scenarios. Experiments with open-source LLMs, such as Mistral (Jiang et al., 2023) and LLaMA (Touvron et al., 2023), will be conducted to assess their suitability for on-premise deployment, ensuring compliance with privacy and regulatory requirements. Collaboration with healthcare professionals will be prioritized to enhance system credibility, refine domain-specific applicability, and improve dataset quality. Additionally, a comprehensive evaluation strategy will be developed to assess fact-correctness and format alignment more effectively.

Limitations

1. It relies on API-based access to GPT-4, which can be expensive and may introduce latency, limiting scalability in real-time clinical use.
2. We consider only two layouts to create the dataset. Datasets with more layouts need to be considered to create future benchmark datasets.

Additionally, our evaluation is based on a synthetic dataset, which may not fully reflect the diversity of real-world genomic reports, potentially affecting generalizability. The system also assumes that all relevant information is within the report images and may struggle with highly specialized content. Finally, relying on AI for critical medical decisions raises ethical concerns, as errors could impact patient care.

Ethics Statement

The GR-DATASET was created synthetically to address the lack of publicly available genomic reports, ensuring no real patient data was used. The dataset simulates key-value structures commonly found in cancer genomic reports. All synthetic data was generated to promote research and development in the medical AI domain, particularly for extracting data from complex documents.

GR-PROMPT and GR-FORMAT techniques are designed to assist researchers, healthcare professionals, and developers in automating the extraction of structured information from genomic reports. However, potential risks include inaccuracies in data extraction due to variability in report formats or domain-specific language that current models may not fully capture. Additionally, biases inherent in the LLM used could affect the extraction accuracy, especially when applied to various genomic reports in the real world. We recommend using these methods cautiously, particularly in clinical settings, and we encourage further validation to ensure their accuracy and reliability in different use cases.

Acknowledgements

We would like to thank the reviewers and Mrs. Settaluri Sravanthi for their valuable feedback.

References

Fda-approved oncology therapies. https://www.oncokb.org/oncology-therapies. Accessed: 2024-09-30.

National cancer institute thesaurus. https://ncithesaurus.nci.nih.gov/ncitbrowser. Accessed: 2024-09-30.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

John Dagdelen, Alexander Dunn, Sanghoon Lee, Nicholas Walker, Andrew S Rosen, Gerbrand Ceder, Kristin A Persson, and Anubhav Jain. 2024. Structured information extraction from scientific text with large language models. *Nature Communications*, 15(1):1418.

Masato Fujitake. 2024. Layoutllm: Large language model instruction tuning for visually rich document understanding. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10219–10224.

Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. 2022. Layoutlmv3: Pre-training for document ai with unified text and image masking. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4083–4091.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and

Seunghyun Park. 2022. Ocr-free document understanding transformer. In *European Conference on Computer Vision*, pages 498–517. Springer.

Chuwei Luo, Yufan Shen, Zhaoqing Zhu, Qi Zheng, Zhi Yu, and Cong Yao. 2024. Layoutllm: Layout instruction tuning with large language models for document understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15630–15640.

David M Miller and Sophia Z Shalhout. 2021. Genetex—a genomics report text mining r package and shiny application designed to capture real-world clinico-genomic data. *JAMIA open*, 4(3):ooab082.

Zhi Rui Tam, Cheng-Kuang Wu, Yi-Lin Tsai, Chieh-Yen Lin, Hung-yi Lee, and Yun-Nung Chen. 2024. Let me speak freely? a study on the impact of format restrictions on performance of large language models. *arXiv preprint arXiv:2408.02442*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Hervé Jégou. 2023. Llama: Open and efficient foundation language models. https://ai.facebook.com/research/publications/llama-open-and-efficient-foundation-language-models/. Accessed: 2024-12-09.

Haolun Wu, Ye Yuan, Liana Mikaelyan, Alexander Meulemans, Xue Liu, James Hensman, and Bhaskar Mitra. 2024. Structured entity extraction using large language models. *arXiv preprint arXiv:2402.04437*.

Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Yuhao Dan, Chenlin Zhao, Guohai Xu, Chenliang Li, Junfeng Tian, et al. 2023. mplug-docowl: Modularized multimodal large language model for document understanding. *arXiv preprint arXiv:2307.02499*.