EHMA 2024

Shaping and managing innovative health ecosystems

Harnessing LLMs for structured clinical data extraction - a tool for informed decision-making in healthcare management

Dr. Mădălin-Marius Margan
MD, PhD, MPH, Asst. Prof.
Victor Babeș University of Medicine and Pharmacy Timișoara Romania

5 - 7 June 2024 - Bucharest, Romania

Politehnica University of Bucharest, Bucharest, Romania

#EHMA2024

Content

- 1. Key benefits of structured data in hospital management
- 2. Uses of LLMs in Healthcare
- 3. LLM Strengths and Limitations
- 4. Usefulness in Data Structuring
- 5. Results of Our own study = Original Article soon in Pre-print:

  A critical assessment of GPT-3.5 and GPT-4 role in Cancer Registry Information: Extraction: evaluating models' performance in original language and after English translation
- 6. Unlocking the potential of LLMs in Healthcare

1. Key benefits of structured data in hospital management

1. Operational Efficiency:
- Streamlined Workflow
- Resource Management

2. Quality and Performance Monitoring:
- Performance Metrics
- Compliance and Reporting

3. Research and Innovation:
- Clinical Research
- Predictive Analytics

4. Patient Engagement and Satisfaction:
- Personalized Care
- Patient Portals

5. Financial Management:
- Billing Accuracy
- Cost Management

6. Enhanced Interoperability:
- Data Exchange
- Integrated Systems

2. General Use Cases for LLMs in Healthcare

1. GENERATION

- ❑ Generating clinical notes and reducing doctor’s documentation time
- ❑ Automating administrative tasks such as medical coding and billing
- ❑ Generating differential diagnosis and improving diagnostic accuracy

2. INFORMATION EXTRACTION

- ✅ Extracting useful information and insights from Electronic Health Records
- ❑ Assisting doctors with literature review and easier access to up-to-date and trustworthy information

2. General Use Cases for LLMs in Healthcare

3. OPERATIONAL EFFICIENCY

- Resource Optimization: By analyzing patterns in hospital admissions, staffing, and resource utilization, LLMs can help optimize schedules and resource allocation.
- Workflow Automation: Automating routine tasks such as scheduling, billing, and reporting

4. PREDICTION & RECOMMENDATION

- Predicting disease progression and identifying high-risk patients
- Recommending personalized treatment plans, interventions, clinical trials, and improving clinical workflows

3. Strenghts vs Limitations of LLMs

Strenghts

- Efficient at synthesizing complex, multimodal clinical data sources
- Can answer complex medical questions with few-to-no examples
- Quite easy to involve in a clinical setting
- Can solve a variety of tasks

Limitations

- **Prone to hallucinations:**

- makes up an answer
- gives a nonsense answer
- gives an answer that conflicts itself

- **Cost:**

- Bilions of parameters

- **Bias:**

- a model trained on biased data may perpetuate stereotypes

Personalized LLMs vs Medical LLMs

Personalized LLMs

- Created to possess extensive knowledge about various subjects.
- Pretrained using diverse data sources across multiple fields.
- Capable of handling both medical and non-medical tasks (may not excel in medical tasks due to limited specialized knowledge).

Medical LLMs

- Designed to have strong domain knowledge of medicine in order to assist in clinical decision-making
- Pretrained on specific medical/biomedical datasets: clinical notes, lab values, etc.
- Likely to perform more strongly on medical tasks (answer complex questions about medical conditions)

+ Fine-Tuning

4. Usefulness in data structuring

- Pattern Recognition and Extraction
- Scalability and Efficiency
- Flexibility across Domains (Specialities)
- Error reduction
- Data Enrichment – risky?
- Transparency and Explainability
- Customization and Fine-Tuning
- Cost-Effectiveness

4. Study

Other Papers on the subject

GPT extraction Performance

A comparative study of zero-shot inference with large language models and supervised modeling in breast cancer pathology classification

Madhumita Sushil, PhD,[1,*] Travis Zack, MD, PhD,[1,2,*] Divneet Mandair, MD,[1,2,*] Zhiwei Zheng, MEng,[3,#] Ahmed Wali, MEng,[3,#] Yan-Ning Yu, MEng,[3,#] Yuwei Quan, MEng,[3,#] and Atul J. Butte, MD, PhD[1,2,4,5]

▸ Author information ▸ Copyright and License information PMC Disclaimer

"The GPT-4 model shows promising capabilities for clinical information extraction, potentially reducing the burden of large-scale data labeling while maintaining high performance in classification tasks."

Parameter	Macro F1-score of GPT in structured data extraction
Histological Type	0.82
ER	0.95
HER-2	0.88

GPT extraction Performance

Journal of Pathology
J Pathol March 2024; **262:** 310–319
Published online 14 December 2023 in Wiley Online Library
(wileyonlinelibrary.com) **DOI:** 10.1002/path.6232

ORIGINAL ARTICLE

Extracting structured information from unstructured histopathology reports using generative pre-trained transformer 4 (GPT-4)

Daniel Truhn[1], Chiara ML Loeffler[2,3,4], Gustav Müller-Franzes[1], Sven Nebelung[1], Katherine J Hewitt[2,4], Sebastian Brandner[5], Keno K Bressem[6], Sebastian Foersch[7] and Jakob Nikolas Kather[2,3,8,9*]

1 *Department of Diagnostic and Interventional Radiology, University Hospital RWTH Aachen, Aachen, Germany*
2 *Else Kroener Fresenius Center for Digital Health, Technical University Dresden, Dresden, Germany*
3 *Department of Medicine I, University Hospital Dresden, Dresden, Germany*
4 *Department of Medicine III, University Hospital RWTH Aachen, Aachen, Germany*
5 *Department of Neurosurgery, University Hospital Erlangen, Erlangen, Germany*
6 *Department of Radiology, Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Berlin, Germany*
7 *Institute of Pathology, University Medical Center Mainz, Mainz, Germany*
8 *Medical Oncology, National Center for Tumor Diseases (NCT), University Hospital Heidelberg, Heidelberg, Germany*
9 *Pathology and Data Analytics, Leeds Institute of Medical Research at St James's, University of Leeds, Leeds, UK*

"We found a high concordance between LLM-generated structured data and human-generated structured data. Consequently, LLMs could potentially be employed routinely to extract ground truth data for machine learning from unstructured pathology reports in the future."

Parameter	Correct extraction of value from digital text after OCR
T	99/100
N	95/100
M	94/100
Ki-67	100/100
Total number of lymph nodes positive	99/100

GPT extraction Performance

A critical assessment of using ChatGPT for extracting structured data from clinical notes

Jingwei Huang, Donghan M. Yang, Ruichen Rong, Kuroush Nezafati, Colin Treager, Zhikai Chi, Shidan Wang, Xian Cheng, Yujia Guo, Laura J. Klesse, Guanghua Xiao, Eric D. Peterson, Xiaowei Zhan & Yang Xie

npj Digital Medicine **7**, Article number: 106 (2024) | Cite this article

"ChatGPT and other large language models present a powerful tool for efficiently and accurately extracting structured data from clinical notes, offering significant advantages over traditional NLP methods in terms of both performance and cost-effectiveness."

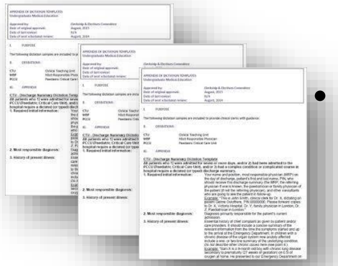
Parameter	F1-score of GPT in structured data extraction
Tumour Stage	0.76
Histological diagnosis	0.99
Primary tumor features (pT)	0.91

Rationale behind the study

- Dilemma: **identify all cases of TNBC** in the Oncology Department
- There was no possibility to automatically extract the relevant cohort from the EHR (there were no filters)
- Even after individual extraction of PDFs for relevant cases, we still had to curate more than 70 cases manually to extract thousands of parameters = useful structured data

=> Necessity to adopt a more feasible implementation in the future

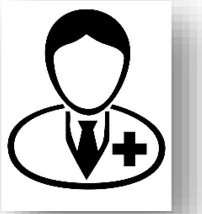
Methodology



PDF extraction + OCR

- A total 769 discharge notes were extracted from the EHR
- An OCR algorithm was used for text extraction from PDFs discharge notes

Creation of the truth dataset



- Manual data curation of same specific by two physicians and two medical students
- A Gold standard dataset was created for comparison

1

2

3

4

Prompt Engineering + LLM calling + Structuring of output

- Personal information was removed, relevant clinical parts were sent to OpenAI API using a Python script
- Output for each case was structured in JSON format => all outputs were transformed in a dataframe
- A specific prompt was composed respecting some literature recommendations

TOTAL COST for AI use: 120$

Calculation of F1-scores

- matplotlib and sklearn.metrics Python libraries were used

Prompt used for data extraction

```
"content": (
    "You are a highly skilled oncologist and clinical researcher. "
    "Analyze the entire text carefully and answer each of the asked questions in a consistent format. "
    "Stick closely to the instructions provided and ensure uniformity in responses across different reports. "
    "Analyze the entire text and answer each of the asked questions in a short format. "
    "(the output should be in the 'key: value' format for each case). "
    "Each answer should be on a separate lane, preceded by the prefix of the question given. "
    "When answering to icd-10 here are the possible answers: C50.011, C50.012, C50.019, "
    "C50.021, C50.022, C50.029, C50.111, C50.112, C50.119, C50.121, C50.122, C50.129, "
    "C50.211, C50.212, C50.219, C50.221, C50.222, C50.229, C50.311, C50.312, C50.319, "
    "C50.321, C50.322, C50.329, C50.411, C50.412, C50.419, C50.421, C50.422, C50.429, "
    "C50.511, C50.512, C50.519, C50.521, C50.522, C50.529, C50.611, C50.612, C50.619, "
    "C50.621, C50.622, C50.629, C50.811, C50.812, C50.819, C50.821, C50.822, C50.829, "
    "C50.911, C50.912, C50.919, C50.921, C50.922, C50.929. "
    "When providing answer to stage of cancer according to AJCC8 choose between the following: "
    "4 3C 3B 3A 2B 2A 1B 1A 0 99. "
    "When providing answer to histological type please choose between the most common 16 icd-o-3 codes: "
    "8500/3 - Infiltrating duct carcinoma, NOS (Not Otherwise Specified), "
    "8520/3 - Lobular carcinoma, NOS, "
    "8522/3 - Infiltrating duct and lobular carcinoma, "
    "8510/3 - Tubular adenocarcinoma, "
    "8521/3 - Infiltrating duct mixed with other types of carcinoma, "
    "8211/3 - Medullary carcinoma with lymphoid stroma, "
    "8230/3 - Cribriform carcinoma, NOS, "
    "8501/3 - Comedocarcinoma, NOS, "
    "8512/3 - Adenoid cystic carcinoma, "
    "8530/3 - Paget disease, NOS, "
    "8503/3 - Intraductal micropapillary carcinoma, "
    "8201/3 - Secretory carcinoma, NOS, "
    "8504/3 - Invasive papillary carcinoma, "
    "8507/3 - Mucinous adenocarcinoma, "
    "8401/3 - Apocrine adenocarcinoma, "
    "8524/3 - Signet ring cell carcinoma. "
    "When providing answer to biomarkers(ER, PR, Ki-67) always use this exact format 'give prefix: Integer(0-100)' "
    "when present, if not mentioned : Not found"
```

Results – F1 scores for:

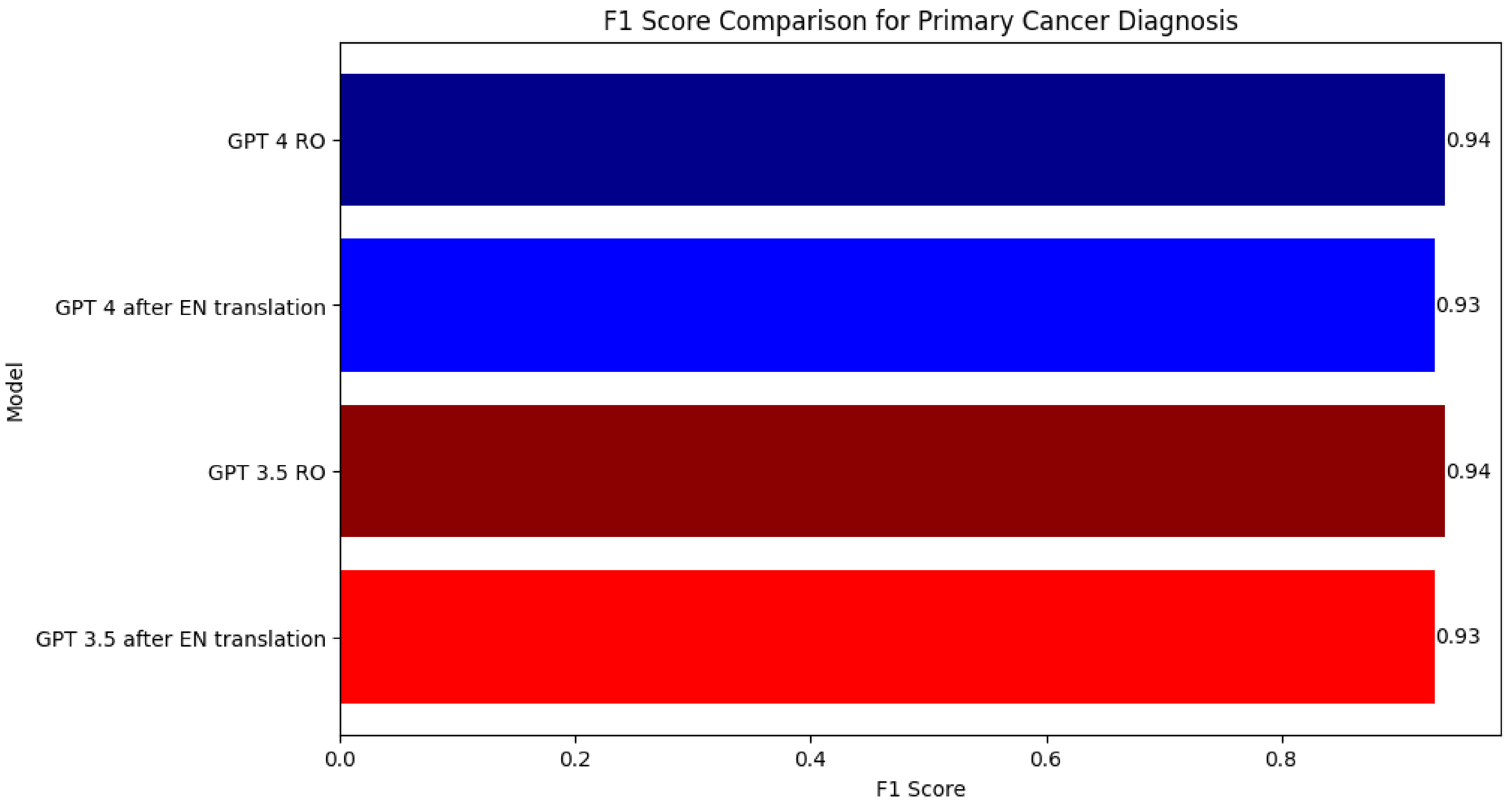
GPT 3.5 on Romanian text (No translation)

GPT 3.5 after English translation

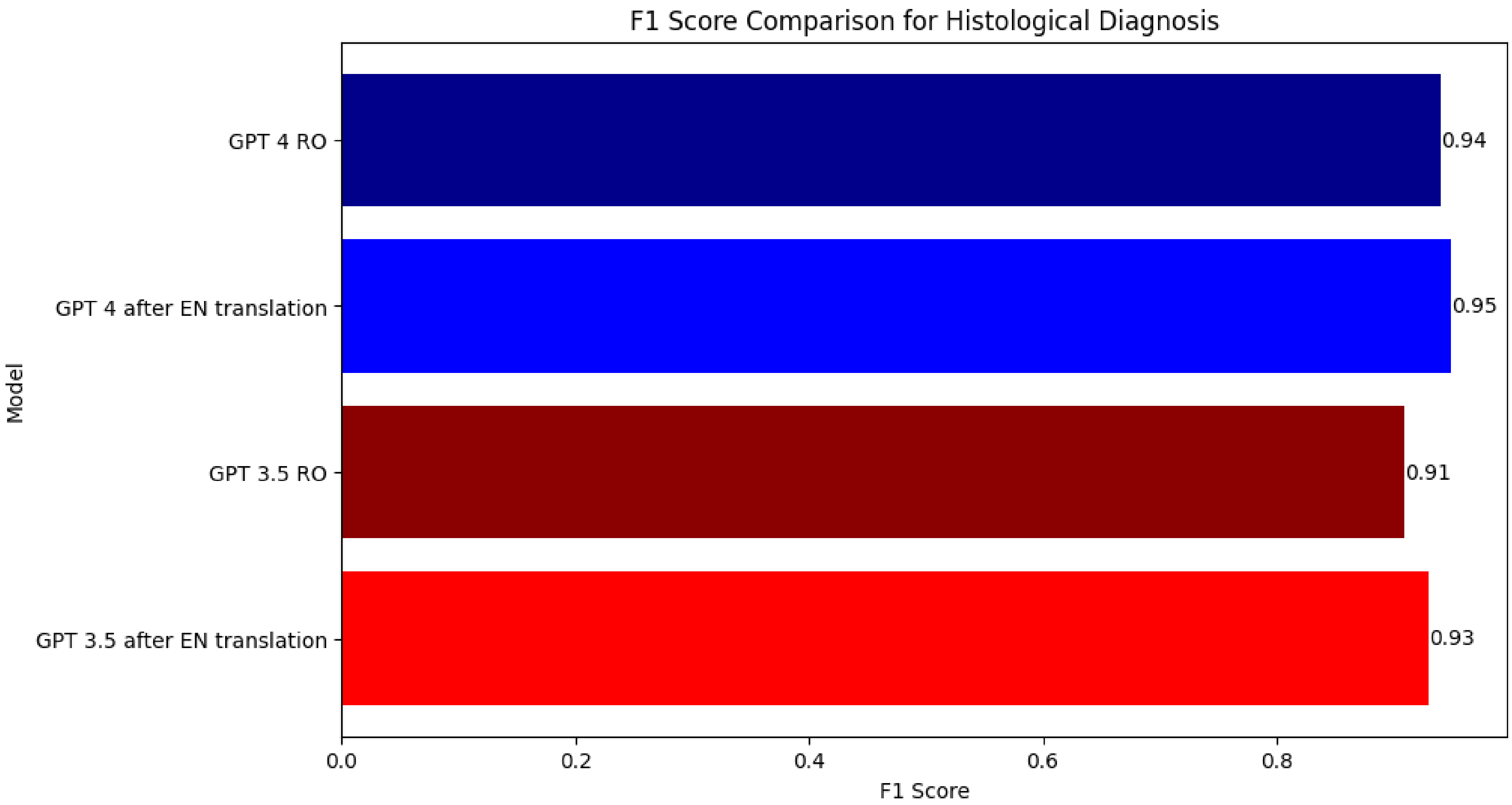
GPT 4 on Romanian text (No translation)

GPT 4 after English translation

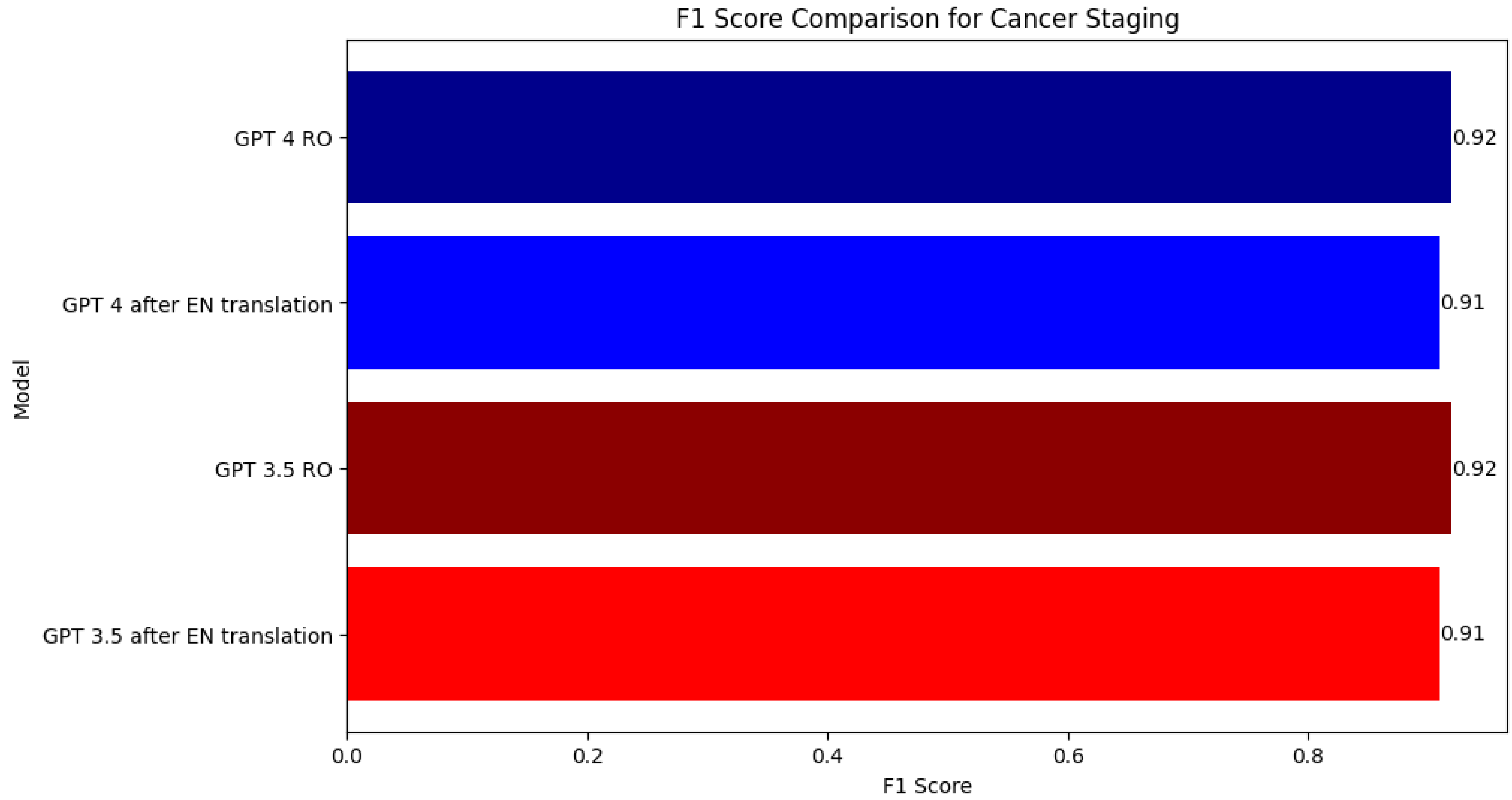
Primary cancer diagnosis (ICD-10 code)



Histological diagnosis (ICD-O-3 code)

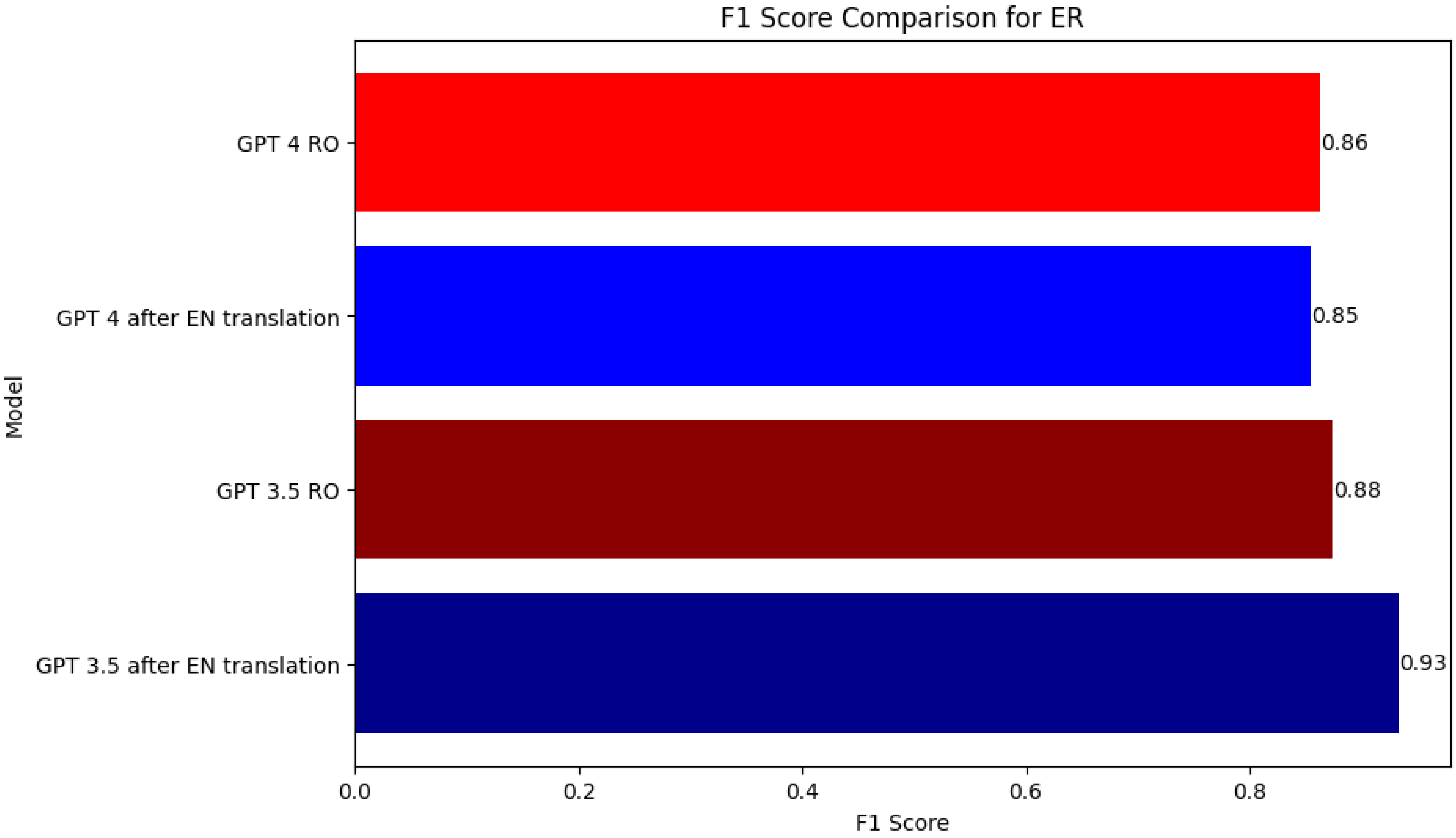


Cancer Staging



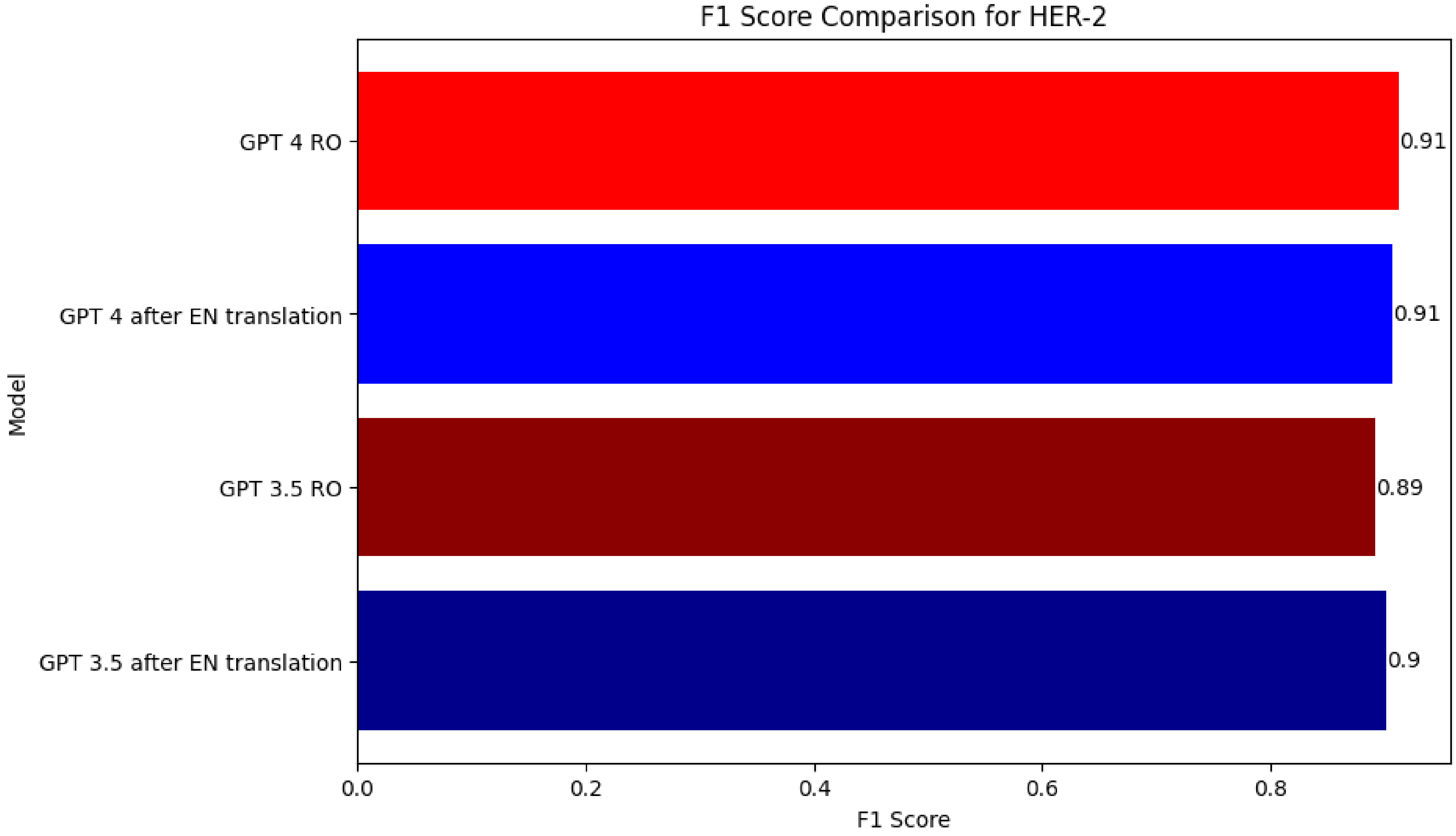
ER

- ✓ GPT 3.5 extracts the first value present in the text (usually the correct one)
- ✓ GPT 4 extracts subsequent values (ingests more context, but in this particular case is not useful)
- ✓ Sometimes referred as HR (combined result with PR)



HER-2

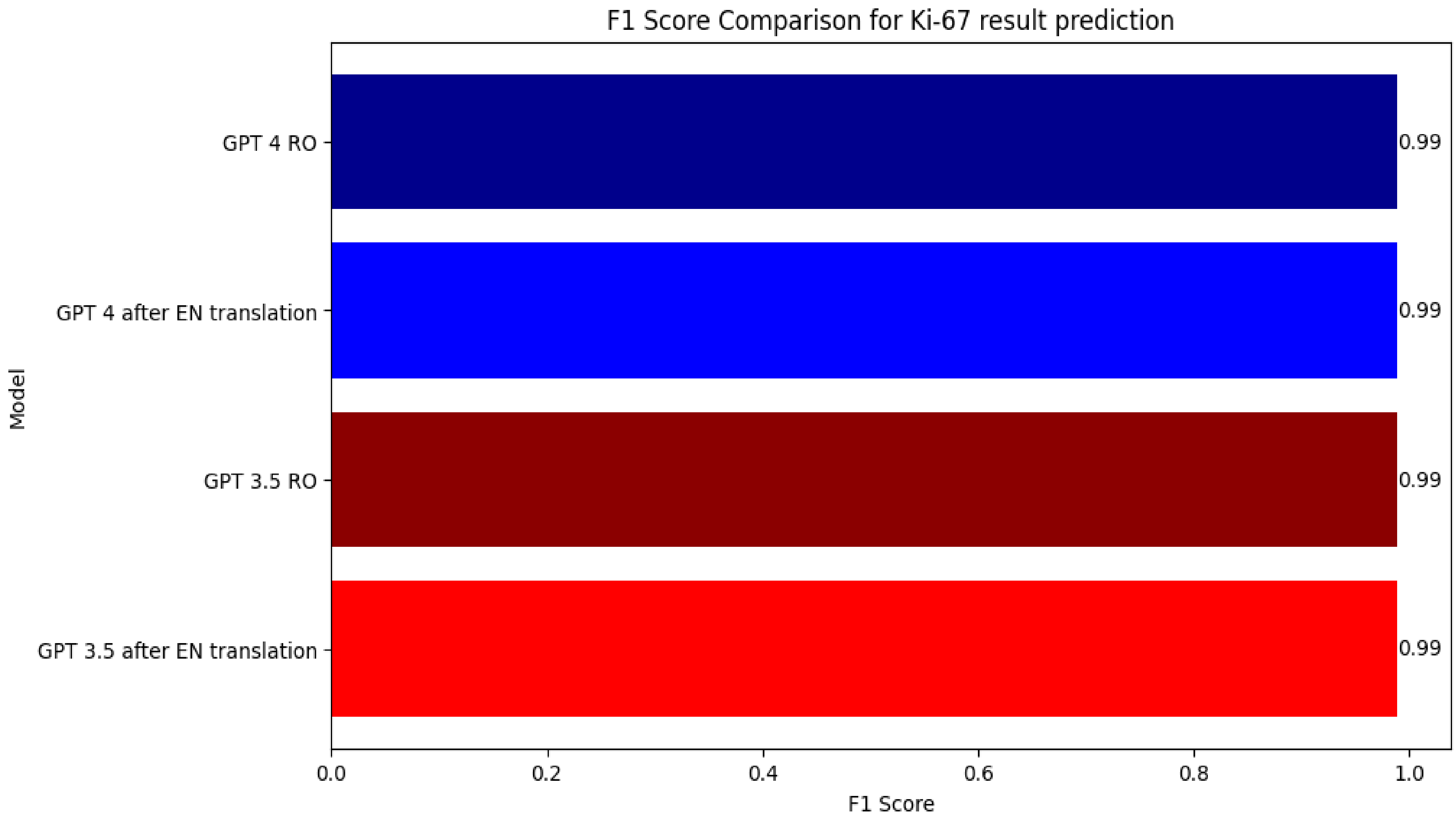
✓ Values such as HER2 appear in different ways:
HER-2, HER 2, HER-2/neu, ERBB2, or c-erbB22
negative, HER2 neg , HER2 – (which can mean not performed or negative) which confuses the algorithm



Ki-67

✓ The "vectorial image" of Ki-67 is very unique

=> Extraction of value is almost perfect



Results: Remarkable Performance in Cancer Data Extraction

1 Histological Diagnosis ✓

2 Cancer Staging ✓

3 Numerical Variables ✓

4 Binary Classification

Adjuvant Therapy versus Absence of Adjuvant Therapy results were paradoxically lower.

Still waiting for results because a new "better" prompt was implemented.

Extraction quality was affected due to lack of reporting standardization

✓ **Inconsistent Terminology Usage**:
- Different physicians may use various abbreviations or terms for the same biomarker or medical condition.
- *Eg: Tumor markers like "HER2" may be referred to differently, such as "HER-2/neu," "erbB-2," or "CD340."*

✓ **Differences in Measurement Units:**
- Lack of standard units for reporting biomarker levels can lead to confusion and misinterpretation of data.
- *Eg: Reporting levels of a biomarker in ng/mL vs. pmol/L without conversion guidelines.*

✓ **Diverse Data Formats**:
- Electronic health records (EHRs) and laboratory reports may use different data formats, complicating data integration and analysis.
- Eg: Free-text vs. structured data entries for the same type of information.

6. Unlocking the Potential of LLMs in Healthcare

both at Institutional Level and for Public Health actors

Improved Data Capture

LLMs can enhance the capture of critical clinical data from unstructured sources, enabling more comprehensive and accurate data-driven decision-making.

Accelerated Research

The efficient extraction of data from cancer registries can significantly benefit medical research, leading to faster identification of relevant patient cohorts and exploration of new treatment options.

Improved Patient Care

By automating data extraction and analysis, LLMs can help healthcare institutions optimize resources, enhance the quality of care, and ultimately improve patient satisfaction.

Driving Innovation

The successful integration of LLMs in healthcare can pave the way for further advancements and the development of more reliable and robust clinical information extraction tools.

Other use cases in hospital operations

- ✓ Identifying the most frequent surgery complications
- ✓ Analyzing adverse events related to medication
- ✓ Defining patient addressability
- ✓ Categorizing patient cohorts
- ✓ Accelerating clinical trial enrollment
- ✓ Examining social aspects of care
- ✓ Conducting public health research by combining multiple data sources
- ✓ Providing reliable information needed for hospital accreditation
- ✓ Calculating average procedure times to optimize different parameters
- ✓ Generating surgical procedure recommendations

Advantage: Discover the problems in initial data input and try to improve it

- ✓ In general: *“When human is confused, machine is confused”*

- ✓ Change in paradigm:

“Physician should write the clinical notes considering that other physicians, as well as AI algorithms, should be able to clearly understand them.
We are not alone anymore when it comes to clinical reasoning.”

Addendum

Privacy concerns

Two large language models, GPT-3.5 and GPT-4, were queried using HIPAA-compliant Azure OpenAI services.

No data was permanently transferred to or stored by either OpenAI or Microsoft, consistent with previous protocols.

A single prompt was used to request all classification labels for each discharge note.

Model outputs, initially in JSON format, were post-processed into Python dataframes for automatic evaluation.

Implications and Future Directions

1

Enhancing Decision-Making

The findings of this study demonstrate the potential of LLMs in extracting consistent structured clinical data, which can empower hospital managers to make more informed decisions, optimize resources, and improve the quality of care.

2

Advancing Medical Research

The efficient extraction of data for disease registries using LLMs can significantly benefit medical research, enabling researchers to quickly identify relevant patient cohorts and explore new treatment options.

3

Driving Innovation

The successful integration of LLMs in healthcare can pave the way for further advancements and the development of more reliable and robust clinical information extraction tools, ultimately improving patient outcomes.

EHMA 2024

Shaping and managing
innovative health ecosystems

Thank you!

Dr. Mădălin-Marius Margan
MD, PhD, MPH, Asst. Prof.
Victor Babeș University of Medicine and Pharmacy Timișoara Romania

E-mail: **margan.madalin@umft.ro**
LinkedIn: **madalinmargan**