Poster Code: MSR64

Using Large Language Models to Extract PD-L1 Testing Details from Electronic Health Records

Aaron B. Cohen, MD, MSCE[1,2]; Michael Waskom, PhD[1]; Blythe Adamson, PhD, MPH[1]; Jonathan Kelly, MEng[1]; Guy Amster, PhD[1]
[1] Flatiron Health, New York, NY; [2] NYU Langone School of Medicine, New York, NY

Background

- The suitability of artificial intelligence (AI) and large language models (LLMs) to assist in curating real-world data (RWD) from electronic health records (EHR) for research has not been extensively evaluated.
- PD-L1 biomarker testing guides cancer treatment decisions. However, results:
  - are hard to access because lab reports are unstructured and require clinical expertise to interpret.
  - vary by cancer type, documentation pattern, and year the test occurred
- This study explored the ability of LLMs to rapidly identify PD-L1 biomarker details in the EHR and the impact of fine-tuning on results.

Methods

- **Data source:** The US nationwide Flatiron Health EHR-derived de-identified database, comprising patient-level structured and unstructured data,[1,2] originating from ~280 cancer clinics (~800 sites of care), majority from community oncology settings.
- **Cohort:** Patients diagnosed with one of 15 cancers after 1/1/2011
- **Primary Outcome:** PD-L1 biomarker testing details
- **Statistical Methods:** Applied open-source LLMs (Llama-2-7B and Mistral-v0.1-7B)[3,4] to extract seven biomarker details relating to PD-L1 testing:
  - Collection/Receipt/Report Date, Cell type, Percent staining, Combined positive score, and Staining intensity.
  - Two approaches: "zero-shot" experiments (no fine-tuning) exploring a range of prompts and fine-tuning on manually-curated answers from 500/1000/1500 documents.
  - Validation: (1) Used 250 expert human abstracted answers across >15 cancer types; (2) compared performance on percent staining to a deep learning model (LSTM) baseline trained on >10,000 examples.[5]

Results

- LLMs extracted all seven biomarker testing details at once from EHR documents.
- Fine-tuned outputs consistently conformed to desired RWD structure.
- Zero-shot outputs were frequently invalid and exhibited hallucination.
- Fine-tuning performance improved with additional training examples:
  - F1 scores ranged from 0.80–0.95, and date accuracy (within 15 days) ranged from 0.85–0.90.
  - Increasing the number of epochs improved performance with limited training examples, but the effect diminished quickly with moderately more training examples.
- Fine-tuned LLMs exceeded performance of deep learning model baseline ($\Delta F1 = 0.05$) despite significant difference in training data.

Fine-tuned LLMs accurately extracted complex biomarker testing details and results from unstructured clinical documents

Scan for abstract and digital poster

Acknowledgments: Darren Johnson, PhD (Flatiron Health, Inc.) provided medical writing support and Madeline Morenberg (Flatiron Health, Inc.) provided design support. Data first presented at ISPOR 2024 in Atlanta, GA on May 7, 2024.

Disclosures: This study was sponsored by Flatiron Health, Inc.—an independent member of the Roche Group. During the study period, ABC, MW, BA, JK, and GA reported employment with Flatiron Health, Inc. and stock ownership in Roche.

Author Contact Info: Aaron B. Cohen, MD, MSCE acohen@flatiron.com

Limitations

- Results may not translate to other biomarkers, and specifically ones that are not standard of care.
- More work is required to see whether fine-tuning on a range of clinical tasks would lead to improved performance.

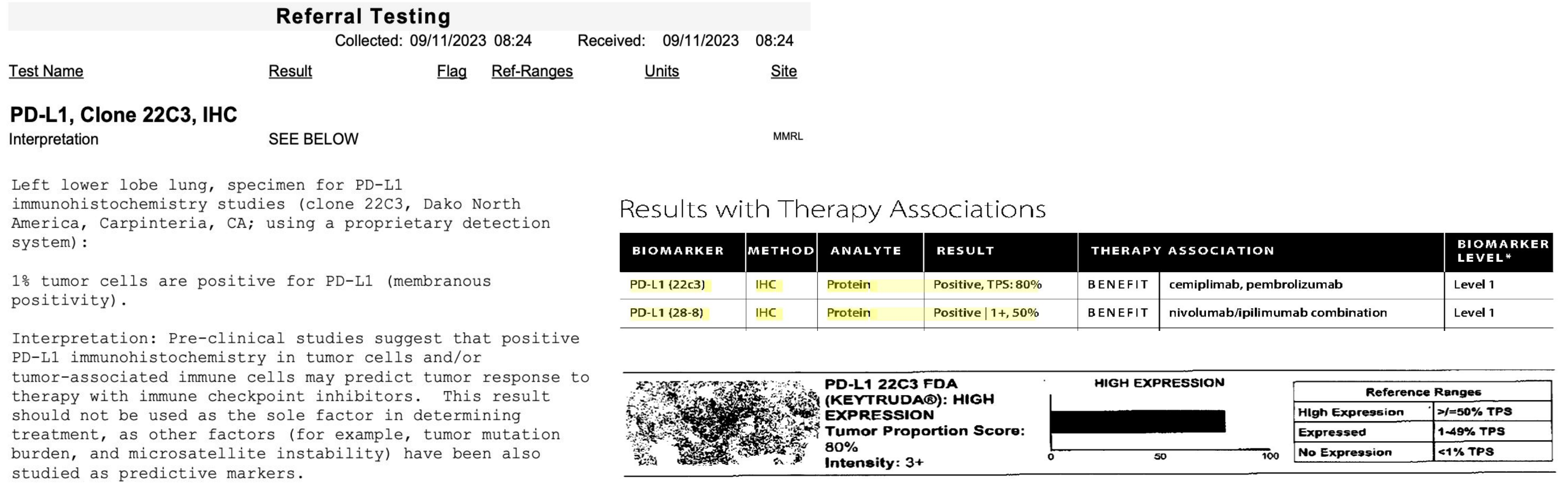
Conclusions

- LLMs, fine-tuned with high-quality labeled data, accurately extracted complex PD-L1 test details from the EHR despite considerable variability in cancer type, documentation, and time.
- Zero-shot prompt extraction not as effective at model scale examined.
- Validation required access to high-quality data labeled by experts with access to the source EHR.

References

1. Ma X et al. MedRxiv. 2023. doi:10.1101/2020.03.16.20037143
2. Birnbaum B et al. arXiv. 2020. doi:10.48550/arxiv.2001.09765
3. Touvron H et al. arXiv. 2023. doi:10.48550/arXiv.2307.09288
4. Jiang AQ et al. arXiv. 2023. doi:10.48550/arXiv.2310.06825
5. Adamson B et al. Front Pharmacol. 2023. doi: 10.3389/fphar.2023.1180962

Figure 1.



Referral Testing

Collected: 09/11/2023 08:24 Received: 09/11/2023 08:24

Test Name	Result	Flag	Ref-Ranges	Units	Site
PD-L1, Clone 22C3, IHC					
Interpretation	SEE BELOW				MMRL

Left lower lobe lung, specimen for PD-L1 immunohistochemistry studies (clone 22C3, Dako North America, Carpinteria, CA; using a proprietary detection system):

1% tumor cells are positive for PD-L1 (membranous positivity).

Interpretation: Pre-clinical studies suggest that positive PD-L1 immunohistochemistry in tumor cells and/or tumor-associated immune cells may predict tumor response to therapy with immune checkpoint inhibitors. This result should not be used as the sole factor in determining treatment, as other factors (for example, tumor mutation burden, and microsatellite instability) have been also studied as predictive markers.

Results with Therapy Associations

BIOMARKER	METHOD	ANALYTE	RESULT	THERAPY ASSOCIATION		BIOMARKER LEVEL*
PD-L1 (22c3)	IHC	Protein	Positive, TPS: 80%	BENEFIT	cemiplimab, pembrolizumab	Level 1
PD-L1 (28-8)	IHC	Protein	Positive | 1+, 50%	BENEFIT	nivolumab/ipilimumab combination	Level 1

PD-L1 22C3 FDA (KEYTRUDA®): HIGH EXPRESSION
Tumor Proportion Score: 80%
Intensity: 3+

HIGH EXPRESSION

Reference Ranges	
High Expression	>/=50% TPS
Expressed	1-49% TPS
No Expression	<1% TPS

ZERO SHOT **FINE TUNED**

Prompt

```
Using the provided document, extract the results of the patient's PD-L1 test.
Return the results in a JSON structure with the following keys:
  - "percent_staining"  # may also be referred to as tumor proportion (TPS)
  - "staining_intensity"
  - "expression_level"
  - "interpretation"  # may be positive or negative/not detected
Only include information that can be found in the document.
```

Output (ZERO SHOT)

```
Sure! Here is the result of the
patient's PD-L1 test in a JSON
structure:
{
"PD-L1": {
"Tumoral Proportion Score (TPS)": 80,
"Clone 22C3": {
"Membranous pattern of expression in
80% of tumor cells": true
}
}
}
Note that the PD-L1 test results are
only available for the tumor sample
(C23-006350). The other sample
(C23-006349) does not have PD-L1 test
results.</s>
```

Output (FINE TUNED)

```
{
    “percent_staining”: “80”,
    “staining_intensity”: “Not Specified”,
    “expression_level”: “High”,
    “interpretation”: “Negative”,
}
```

Figure 2.

Ability to parse response into json, and into valid json

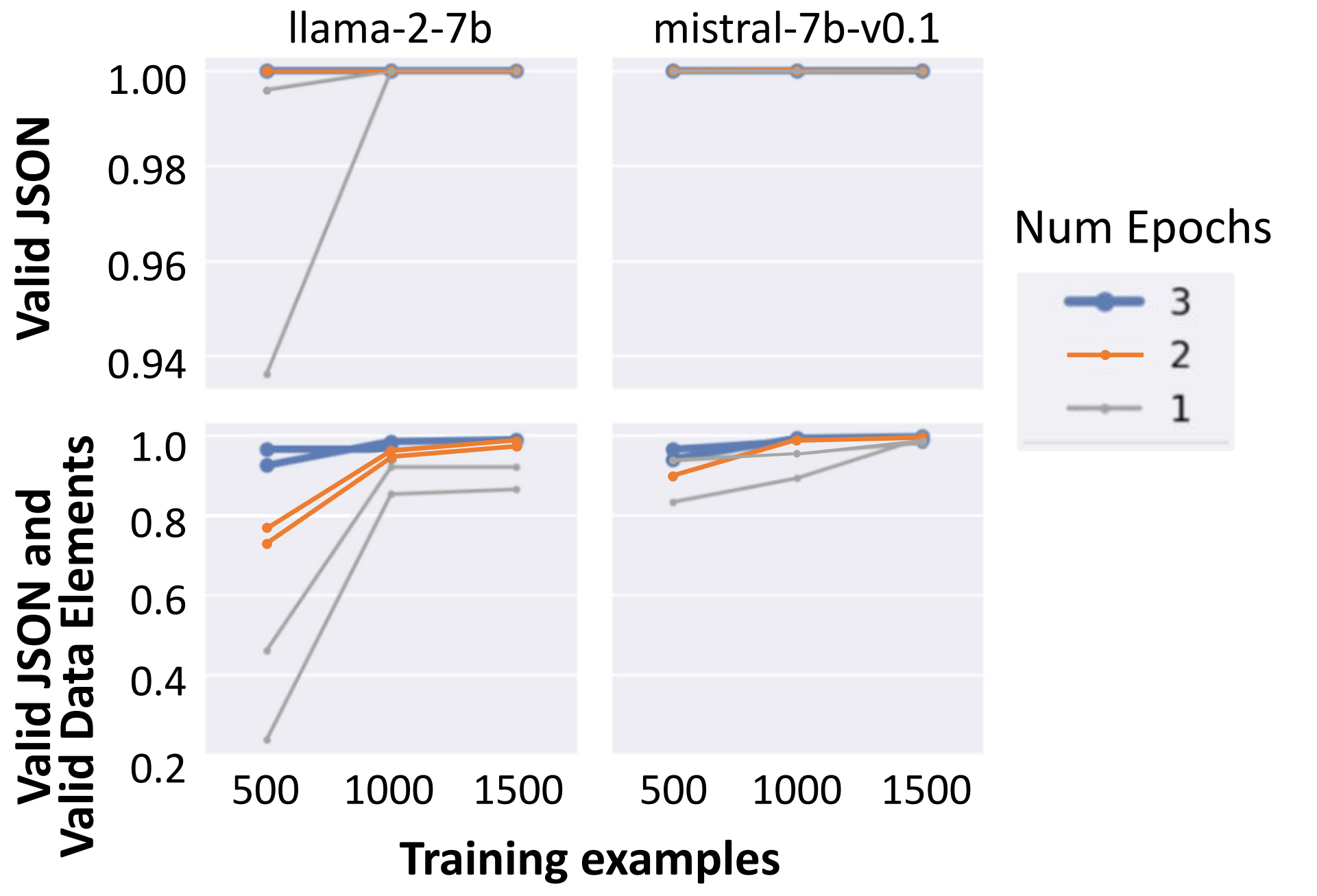
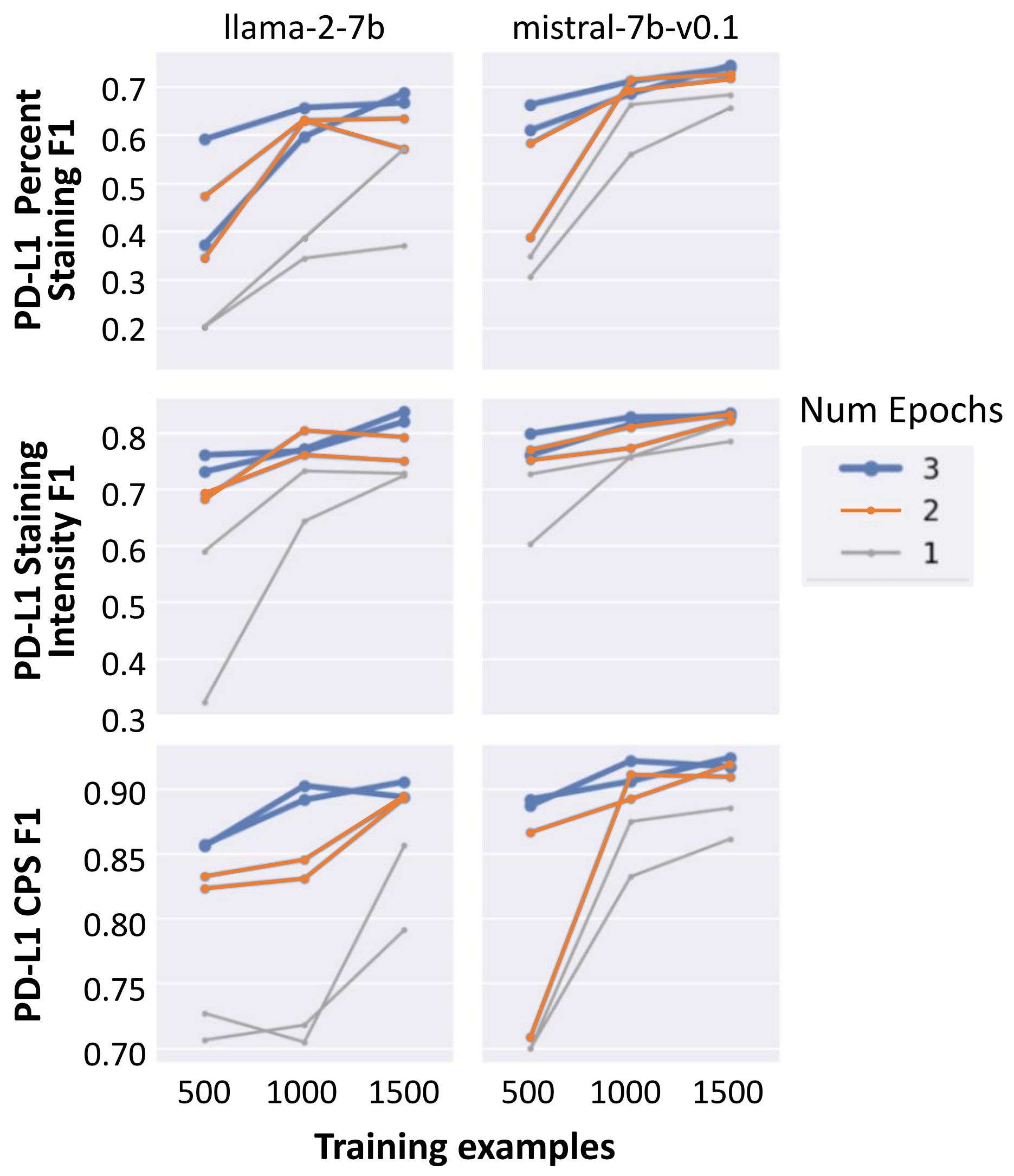


Figure 3.

F1 Score for different extracted variables



llama-2-7b
mistral-7b-v0.1
PD-L1 Percent Staining F1
PD-L1 Staining Intensity F1
PD-L1 CPS F1
Training examples
Num Epochs