Chart Abstraction using Llama 3.2 Lightweight Models

Hyunjoon Lee

1. Overview

1. Medical informatics research on suicide often utilizes extensive electronic health record (EHR) data.

2. Unfortunately, diagnostic codes for suicide attempts (SA) are frequently under-reported and under-coded.

3. Recent studies have turned to natural language processing (NLP) methods to identify SA events within clinical notes.

4. Nevertheless, employing NLP in clinical research typically relies on highly trained experts, as current state-of-the-art methods involve complex text-mining techniques due to the **(1) vast number of texts (hundreds of million)** and **(2) security concerns.**

5. **This reliance creates a bottleneck for tasks such as chart abstraction from clinical notes.**

6. *Lightweight LLMs like Llama 3.2 1B and 3B models offer a solution by enabling faster and more cost-effective computation, and by running locally, they mitigate security risks.*

2. Presentation materials

1. **Expert annotated texts**
2. **Prompt engineering**
3. **Raw model inference**
4. **Fine-tune**
5. **Fine-tuned model inference**
6. **Performance evaluation**

1. Expert annotated texts

Clinical note text	True Label	Llama Label
Pt is a 19 year old …	1 (suicide attempt)	**?**
Pt was transferred from …	0 (x suicide attempt)	**?**
Total: 342 labeled notes on suicide		

2. Prompt engineering (JSON output)

f"""[INST]
Task: You are a clinical expert analyzing notes for evidence of suicide attempt.

Examples of strong evidence of suicide attempt:
{strong_evidence_example}

Example of weak evidence of suicide attempt:
{weak_evidence_example}

Analyze the following clinical note and provide EXACTLY two outputs in the specified format:
Clinical Note: "{note_text}"

Rules:
1. CLASSIFICATION must be a single number:
    0: Does not indicate suicide attempt.
    1: Indicates suicide attempt.

YOU MUST FORMAT YOUR RESPONSE INTO a JSON FORMAT EXACTLY AS FOLLOWS:

{{"person_id": {person_id},
  "note_date": "{note_date}",
  ~~“evidence”: <Write your evidence analysis here.>~~
  "classification": <Single number only: must be 0 or 1>}}

Again, very important, please provide the JSON string with all inner double quotes properly escaped.
[/INST]"""

3. **Model inference**
    - See Jupyter Notebook

4. **Fine-tune**
    - See Jupyter Notebook

5. Performance evaluation

- The Lightweight Llama 3.2 models demonstrated inconsistent reliability in generating JSON format outputs.

- Out of 363 attempts, the Llama 3.2 1B Instruct model failed to produce the correct output **26 times (7%)**, the 3B Instruct model failed **64 times (18%)**, and the fine-tuned 3B model failed **73 times (20%)**.

		Actual	
		0	**1**
Pred	**0**	**TN**	**FN**
	1	**FP**	**TP**

1. **PPV: TP / (TP + FP)**
    - **Among predicted POS, % actually are POS.**

2. **Sensitivity: TP / (TP + FN)**
    - **Among actual POS, % identified correctly.**

Initial Inference

Llama 3.2 1B Instruct			**Llama 3.2 3B Instruct**			**Llama 3.2 3B Fine-tuned**		
	Actual			Actual			Actual	
Pred	**26**	27	Pred	**29**	11	Pred	**25**	6
	114	**170**		94	**165**		101	**158**
PPV: 0.60			**PPV: 0.64**			**PPV: 0.61**		
Sen: 0.86			**Sen: 0.94**			**Sen: 0.96**		
Dropped: 26			**Dropped: 64**			**Dropped: 73**		

- **The model struggled to differentiate between notes indicating strong evidence of <u>suicide attempt</u> from <u>suicidal ideation.</u>**

		suicide attempt	
		0	**1**
Suicidal Ideation	**0**	**72**	**0**
	1	**70**	**200**

Inference without notes indicating suicidal ideation

Llama 3.2 1B Instruct			**Llama 3.2 3B Instruct**			**Llama 3.2 3B Fine-tuned**		
	Actual			Actual			Actual	
Pred	**13**	27	Pred	**20**	11	Pred	**17**	6
	57	**170**		41	**165**		47	**158**
PPV: 0.76			**PPV: 0.80**			**PPV: 0.77**		
Sen: 0.86			**Sen: 0.94**			**Sen: 0.96**		
Dropped: 21			**Dropped: 51**			**Dropped: 60**		

- **Change the prompt.**
  ~~“Example of weak evidence of suicide attempt:”~~
  “Example of strong evidence of suicidal ideation but not suicide attempt:”

Inference using updated prompt

Llama 3.2 1B Instruct			Llama 3.2 3B Instruct			Llama 3.2 3B Fine-tuned		
	Actual			Actual			Actual	
Pred	**33**	32	Pred	**43**	27	Pred	**52**	25
	107	**165**		82	**146**		75	**138**
~~**PPV: 0.60**~~ **PPV: 0.61**			~~**PPV: 0.64**~~ **PPV: 0.64**			~~**PPV: 0.61**~~ **PPV: 0.65**		
~~**Sen: 0.86**~~ **Sen: 0.84**			~~**Sen: 0.94**~~ **Sen: 0.84**			~~**Sen: 0.96**~~ **Sen: 0.85**		
Dropped: 26			**Dropped: 65**			**Dropped: 73**		

GPT4o-mini			Text mining method					
	Actual			Actual				
Pred	**57**	2	Pred	**124**	127			
	83	**185**		18	**73**			
PPV: 0.69			**PPV: 0.80**					
Sen: 0.99			**Sen: 0.365**					
Dropped: 21			**Dropped: 0**					

3. Critical Analysis

1. The performance of lightweight models in chart abstraction seems adequate for non-subtle tasks, but they **exhibit significant deficiencies when applied to more nuanced tasks**.

2. **Fine-tuning these models with clinical notes did not yield any performance improvements**, likely because they have already been trained on publicly available clinical notes datasets.

3. The Lightweight Llama 3.2 models have shown **inconsistent reliability in generating outputs in JSON format**.

4. Should the reliability issues be addressed, lightweight language models could become highly effective tools for conducting chart abstraction on extensive collections of clinical texts locally. This would enable individuals without expertise in natural language processing to conduct chart abstraction, thereby promoting broader access and usability.

5. Future research should focus on evaluating the model's performance in identifying **social determinants of health (SDoH) variables, especially those that are less nuanced**.

4. Resource Links

1. https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/
2. https://github.com/meta-llama/llama-models/tree/main
3. https://github.com/FrostNT1/Llama-3.2-presentation
4. https://colab.research.google.com/drive/1T5-zKWM_5OD21QHwXHiV9ixTRR7k3iB9?usp=sharing
5. https://www.brimanalytics.com/