

Meta AI Llama 3.2 & 3.3 (API)

By Dr. Chris J. Lu
NIH/NLM/LHNCBC/ACIB

April 24, 2025
Uniformity & Conformity Technical Meeting

Slides: <https://wiki.nlm.nih.gov/confluence/display/LHC/LHC+Utilizing+OpenAI+GPT+for+Healthcare+Research>



National Library of Medicine
Lister Hill National Center for Biomedical Communications

Outlines

- Overview of Meta AI Llama
- Use Case Results – Identifying SDOH from Provider's Notes
- Evaluation & Conclusion
- Discussion and Q & A

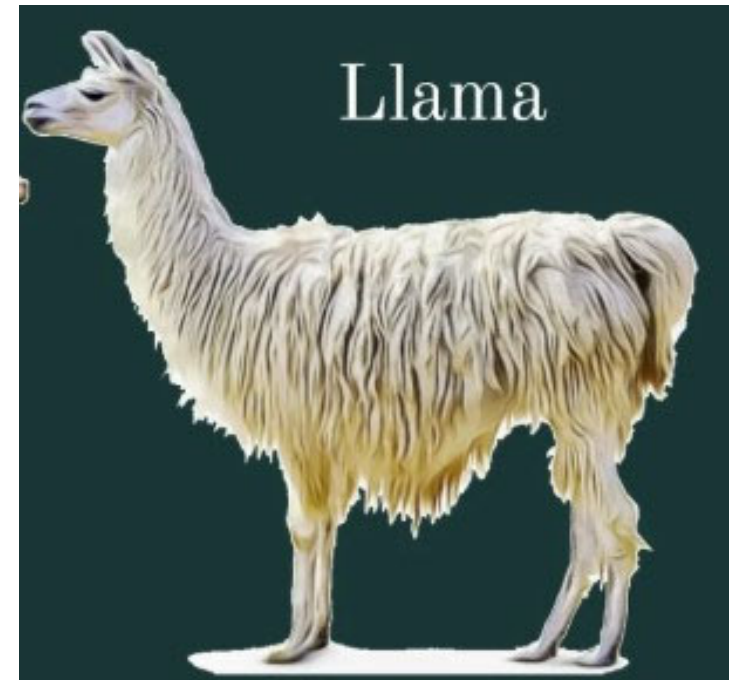


Overview of the LLM Pilot Study

- OpenAI API Access:
 - Requires API key
 - Fee required
 - Account currently on hold (as of 01/27/2025).
- Meta AI Llama API:
 - Run on-premise
 - No usage fee
 - Supports **OpenAI API protocol** (via Meta AI API – post method).
 - Llama 3.2 lightweight 3B model (host: lhcx-gpudev10.nlm.nih.gov).
 - Llama 3.3 70B model (host: lhcx-gpudev20.nlm.nih.gov).

Meta AI Llama

- Llama: Large Language Model by Meta AI
 - a series of open-source AI models (LLMs) developed by Meta AI, designed to be fine-tuned and deployed anywhere.
 - multi-modal: text, images, (audio, video).
 - Meta AI (Meta Platforms, Inc.) is the parent company of Facebook.
 - Releases: 2, 3, 4



National Library of Medicine

Lister Hill National Center for Biomedical Communications

Comparison on Llama 3.2, 3.3 and GPT-4o

Feature	Meta AI Llama 3.2	Meta AI Llama 3.3	OpenAI GPT-4o
Release Date	09.25.2024	12.06.2024	05.13.2024
Model Size	1B, 3B , (11B, 90B)	70B	Estimated 200B+
Context Window	128K tokens	128K tokens	128K tokens
Vocabulary Size	128K	128K	200K (199,997)
Multimodal	Text-only (images)	Text only	Text + audio + image+ video
Type	Chat model	Chat model	Chat model
G-P-T	Yes-Yes-Yes	Yes-Yes-Yes	Yes-Yes-Yes
Multilingual	8 languages	8 languages	77 languages
Knowledge cutoff	12.2023	12.2023	10.2023
Accessibility	Open source	Open source	Proprietary



EHR-01: Information Extraction

Patient Information

- **Patient Name:** John Doe
- **Age:** 45
- **Gender:** Male
- **Date of Birth:** [Insert DOB]
- **Contact Information:** [Insert Phone Number, Address]

Chief Complaint

Patient presents with high stress levels, difficulty accessing fresh food, and concerns about job insecurity.

Medical History

- **Chronic Conditions:** None reported.
- **Previous Surgeries:** None reported.
- **Medications:** Not currently taking any medications.

Social History

- **Living Situation:** Lives alone.
- **Employment:** Works in a low-paying job without health insurance.
- **Diet:** Reports difficulty accessing fresh food due to transportation issues.
- **Support System:** No family support nearby.
- **Lifestyle:** Sedentary lifestyle due to work constraints; struggles with maintaining a healthy diet.

Review of Systems

- **General:** Reports feeling stressed; no weight changes noted.
- **Cardiovascular:** No chest pain, palpitations, or edema.
- **Respiratory:** No shortness of breath or cough.
- **Gastrointestinal:** No significant changes in appetite; reports occasional gastrointestinal discomfort due to stress.
- **Musculoskeletal:** No joint pain or muscle aches reported.
- **Psychiatric:** High levels of stress related to job insecurity and living conditions; no history of mental health disorders reported.

Physical Examination

- **Vital Signs:**
 - Blood Pressure: [Insert BP]
 - Heart Rate: [Insert HR]
 - Respiratory Rate: [Insert RR]
 - Temperature: [Insert Temp]
- **General Appearance:** Appears anxious.
- **Cardiovascular:** Heart sounds normal; no murmurs detected.
- **Respiratory:** Clear upon auscultation; no wheezing or crackles.
- **Abdomen:** Soft, non-tender; normal bowel sounds.
- **Neurological:** Alert and oriented; no focal deficits.

Provider's Notes

John stays at a shelter and works part-times in a low-paying job without health insurance. He mentioned difficulty in accessing fresh food due to transportation issues. Reported high stress related to job insecurity and living conditions. No family support nearby.

Plan

1. **Referral to Social Services:** Connect with local food banks and community programs for assistance.
2. **Mental Health Support:** Encourage John to seek counseling services or support groups to manage stress and improve coping strategies.
3. **Follow-Up Appointment:** Schedule a follow-up in 4–6 weeks to assess progress and re-evaluate social support and stress levels.
4. **Health Education:** Provide resources on nutrition and stress management techniques.
5. **Consider Insurance Options:** Discuss potential avenues for obtaining health insurance coverage, including local programs.

Signature

[Provider's Name]

[Provider's Title]

[Date]



National Library of Medicine

Lister Hill National Center for Biomedical Communications

EHR-01: Provider's Notes to SDOH

Provider's Notes

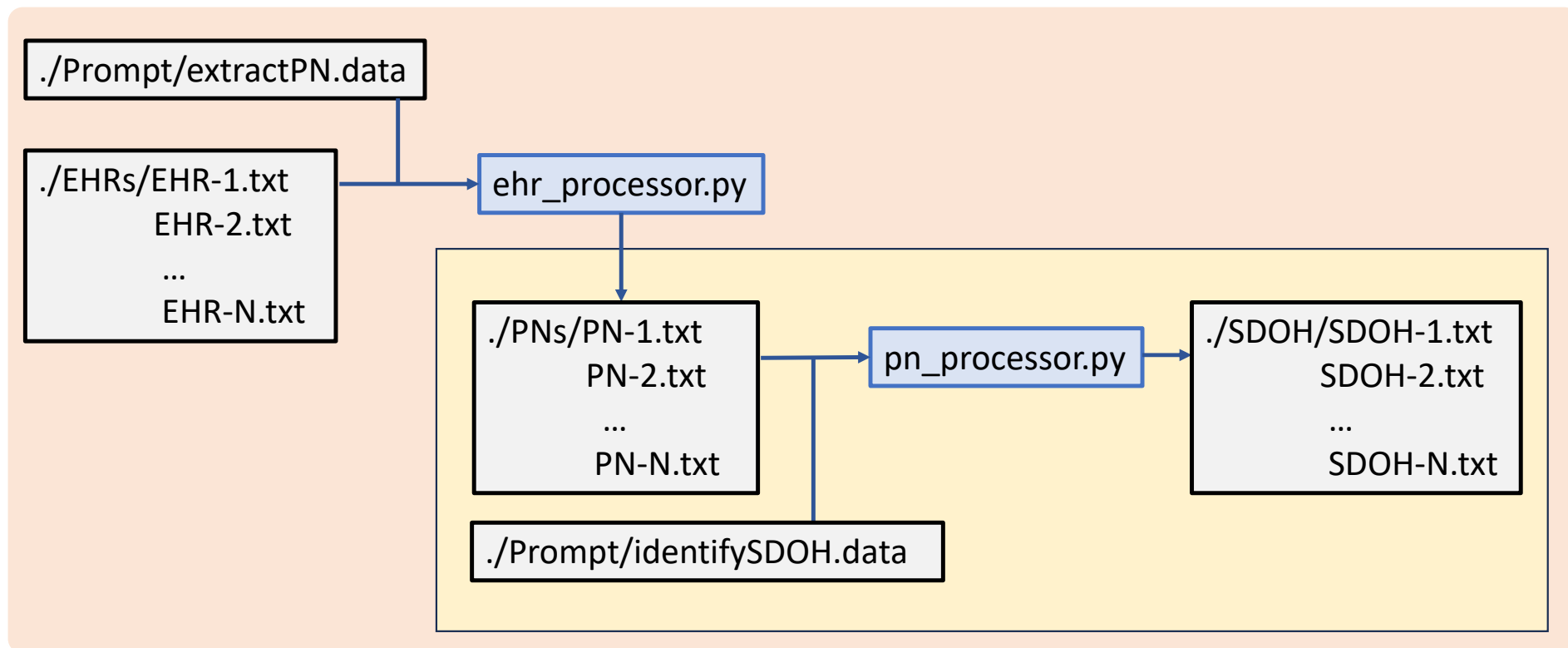
John stays at a shelter and works part-times in a low-paying job without health insurance. He mentioned difficulty in accessing fresh food due to transportation issues. Reported high stress related to job insecurity and living conditions. No family support nearby.



1. **Housing Stability:** stays at a shelter.
2. **Employment and Income:** works part-time in a low-paying job.
3. **Health Insurance Access:** without health insurance.
4. **Food Security:** difficulty in accessing fresh food.
5. **Social Support:** no family support nearby.
6. **Stress and Mental Health:** high stress related to job insecurity and living conditions.



LLM API - Identify SDOH Project



National Library of Medicine

Lister Hill National Center for Biomedical Communications

extractPN.data (prompt)

Please **extract** the provider's notes from the following EHR text.
EHR text: ``{ehr}``

If no provider's notes are found, return "None".

Return only the notes, without any additional information.

NLP capabilities	Requirement
Reading Comprehension	Some – extract
Commonsense Reasoning	Little – if .., return None
World Knowledge	No



identifyPN.data (prompt)

You are a clinical expert.

Please analyze the following Provider's notes text to **identify** Social Determinants of Health (SDOHs) and the associated reason that may affect patient's well-being.

Provider's notes text: "{pn}"

Find the Z-code for each identified SDOH from ICD-10-CM (International Classification of Diseases, 10th Revision, Clinical Modification) coding system.

Please format the response **as a list of JSON object with indent 2 with SDOH, Z-code, reason as the keys.**

Return an empty list if no SDOH is identified.

Please do NOT include extra information or explanation except for the JSON object.

NLP capabilities	Requirement
Reading Comprehension	more – analyze, identify, find, etc.
Commonsense Reasoning	more – SDOH, Z-code, JSON, if .., return None
World Knowledge	Yes – SDOH, Z-code, list, JSON, indent

Results Comparison: Llama 3.2 vs. GPT-4o

API: Llama 3.2 (3B)	Chatbot: Chat GPT-4o
<pre>{ "SDOH": "Unstable Housing", "Z-code": "L00-L09", "reason": "Living in a shelter with limited access to basic necessities" }, { "SDOH": "Food Insecurity", "Z-code": "C02-C08", "reason": "Difficulty accessing fresh food due to transportation issues" }, { "SDOH": "Employment and Income", "Z-code": "P00-P99", "reason": "Low-paying job without health insurance, leading to financial stress" }, { "SDOH": "Social Isolation", "Z-code": "S06-S09", "reason": "No family support nearby, contributing to high stress levels" } }</pre>	<pre>{ "SDOH": "Housing instability (shelter residence)", "Z-code": "Z59.0", "reason": "John stays at a shelter." }, { "SDOH": "Employment and occupational concerns", "Z-code": "Z56.0", "reason": "Works part-time in a low-paying job without health insurance." }, { "SDOH": "Food insecurity", "Z-code": "Z59.4", "reason": "Difficulty accessing fresh food due to transportation issues." }, { "SDOH": "Lack of adequate social support", "Z-code": "Z63.31", <= Absence of family member due to military deployment "reason": "No family support nearby." }, { "SDOH": "Stress related to life circumstances", "Z-code": "Z65.8", "reason": "High stress related to job insecurity and living conditions." } }</pre>



National Library of Medicine

Lister Hill National Center for Biomedical Communications

Results Comparisons: Llama 3.3 vs. GPT-4o

API: Llama 3.3 (70B)	Chatbot: Chat GPT-4o
<pre>{ "SDOH": "Housing instability", "Z-code": "Z59.0", <= Homelessness "reason": "John stays at shelter" }, { "SDOH": "Employment instability", "Z-code": "Z56.81", <= Sexual harassment on the job "reason": "low-paying job without health insurance and job insecurity" }, { "SDOH": "Food insecurity", "Z-code": "Z59.4", <= Lack of adequate food "reason": "difficulty in accessing fresh food due to transportation issues" }, { "SDOH": "Social isolation", "Z-code": "Z91.89", <= Other specified personal risk factors, not elsewhere classified "reason": "No family support nearby" } }</pre>	<pre>{ "SDOH": "Housing instability (shelter residence)", "Z-code": "Z59.0", <= Homelessness "reason": "John stays at a shelter." }, { "SDOH": "Employment and occupational concerns", "Z-code": "Z56.0", <= Unemployment, unspecified "reason": "Works part-time in a low-paying job without health insurance." }, { "SDOH": "Food insecurity", "Z-code": "Z59.4", <= Lack of adequate food "reason": "Difficulty accessing fresh food due to transportation issues." }, { "SDOH": "Lack of adequate social support", "Z-code": "Z63.31", <= Absence of family member due to military deployment "reason": "No family support nearby." }, { "SDOH": "Stress related to life circumstances", "Z-code": "Z65.8", "reason": "High stress related to job insecurity and living conditions." } }</pre>

Performance Comparison on EHR-01

- SDOH

Model	Precision	Recall	F1	Accuracy
Llama 3.2	1.00	0.67	0.80	0.67
Llama 3.3	1.00	0.67	0.80	0.67
GPT-4o	1.00	0.83	0.91	0.83

- Z-code

Model	Precision	Recall	F1	Accuracy
Llama 3.2	0.00	0.00	0.00	0.00
Llama 3.3	0.75	0.60	0.67	0.50
GPT-4o	0.80	0.80	0.80	0.67

Assessment: Llama 3.2 vs. 3.3

- Tested on 69 Python programs.
- Manually rating scale from 0 – 5* based on the responses in the following categories:
 - General Functions.
 - NLP Capabilities.
 - Performance on NLP Applications.
 - Prompt Handling.
 - Input/Output Performance.

*

No	Very Poor	Poor	Fair	Good	Excellent
0	1	2	3	4	5



Llama 3.2 vs. 3.3: General Function Comparison

General Function	Llama 3.2	Llama 3.3
Generative	Yes (5)	Yes (5)
Code - generation, in-filling, efficient	Yes (4)	Yes (4)
Role play (persona)	Yes (3)	Yes (4)
Tone modulation	Yes (4)	Yes (4)
Spelling and grammar	Yes (3)	Yes (4)
Proofread and correct	Yes (3)	Yes (4)
Rewrite	Yes (5)	Yes (5)
Verify	Yes (5)	Yes (5)



Llama 3.2 vs. 3.3: NLP Capabilities Comparison

NLP Capability	Llama 3.2	Llama 3.3
Reading comprehension	Yes (3)	Yes (4)
Commonsense reasoning	Yes (2)	Yes (4)
Word knowledge: in general	Yes (3)	Yes (4)
Word knowledge: ICD-10-CM Z-code	No (0)	Yes (4)
Word knowledge: UMLS concepts (CUI)	No (0)	Yes (4)
Knowledge injection	Yes (4)	Yes (4)



Llama 3.2 vs. 3.3: NLP Applications Comparison

NLP Applications	Llama 3.2	Llama 3.3
Information extraction	Yes (5)	Yes (5)
Information extraction with focus	Yes (3)	Yes (4)
Text summarization	Yes (4)	Yes (5)
Text summarization with focus	Yes (3)	Yes (4)
Sentiment analysis	Yes (5)	Yes (5)
Emotion identification	Yes (4)	Yes (5)
Text classification	Yes (3)	Yes (5)
Information identification	Yes (3)	Yes (5)
Question answering (auto reply)	Yes (3)	Yes (4)
Language translation	Yes (4)	Yes (4)
Language detection	Yes (4)	Yes (4)



Llama 3.2 vs. 3.3: Prompt Handling Comparison

Prompting Handling	Llama 3.2	Llama 3.3
Zero shot prompting	Yes (4)	Yes (4)
Few-shot prompting	Yes (2)	Yes (4)
Chain-of-Thought prompting	Yes (2)	Yes (4)
Multi-task within one prompt	Yes (2)	Yes (4)
Batch jobs prompt utilization (APIs)	Yes (5)	Yes (5)
Context retention (within one session)	Yes (5)	Yes (5)

Llama 3.2 vs. 3.3: Input/Output Comparison

Input & Output Performance	Llama 3.2	Llama 3.3
Input – safeguard	Yes (4)	Yes (4)
Input - prompt format	Yes (3)	Yes (3)
Output – safeguard	Yes (4)	Yes (4)
Output - max. tokens	Yes (5)	Yes (5)
Output - format conversion	Yes (3)	Yes (3)
Output - JSON format	Yes (3)	Yes (4)
Output – table	Yes (3)	Yes (3)
Output - consistency	Yes (4)	Yes (5)
Output - concise (overly verbose Issue)	Yes (2)	Yes (5)
Output- mitigate hallucinations	Yes (3)	Yes (4)



Conclusion

- Meta AI Llama 3.3 (70B) demonstrates advanced NLP capabilities, making it a valuable tool for healthcare research at LHNCBC.
- Meta AI Llama 3.2 (3B) offers more basic NLP functionalities and faster speed, and it is well-suited for simpler NLP tasks.
- Both models provide the benefit of no additional cost, making them accessible and cost-effective options for research use.



Discussion and Q & A

- Current Experiment Status:
 - Experimenting on a mid-size dataset (13K records)
 - Using the Mega prompt (large and highly detailed).
- Next Steps:
 - Focus on large-scale healthcare research.
- Question:
 - Are we ready to provide LLM resources?