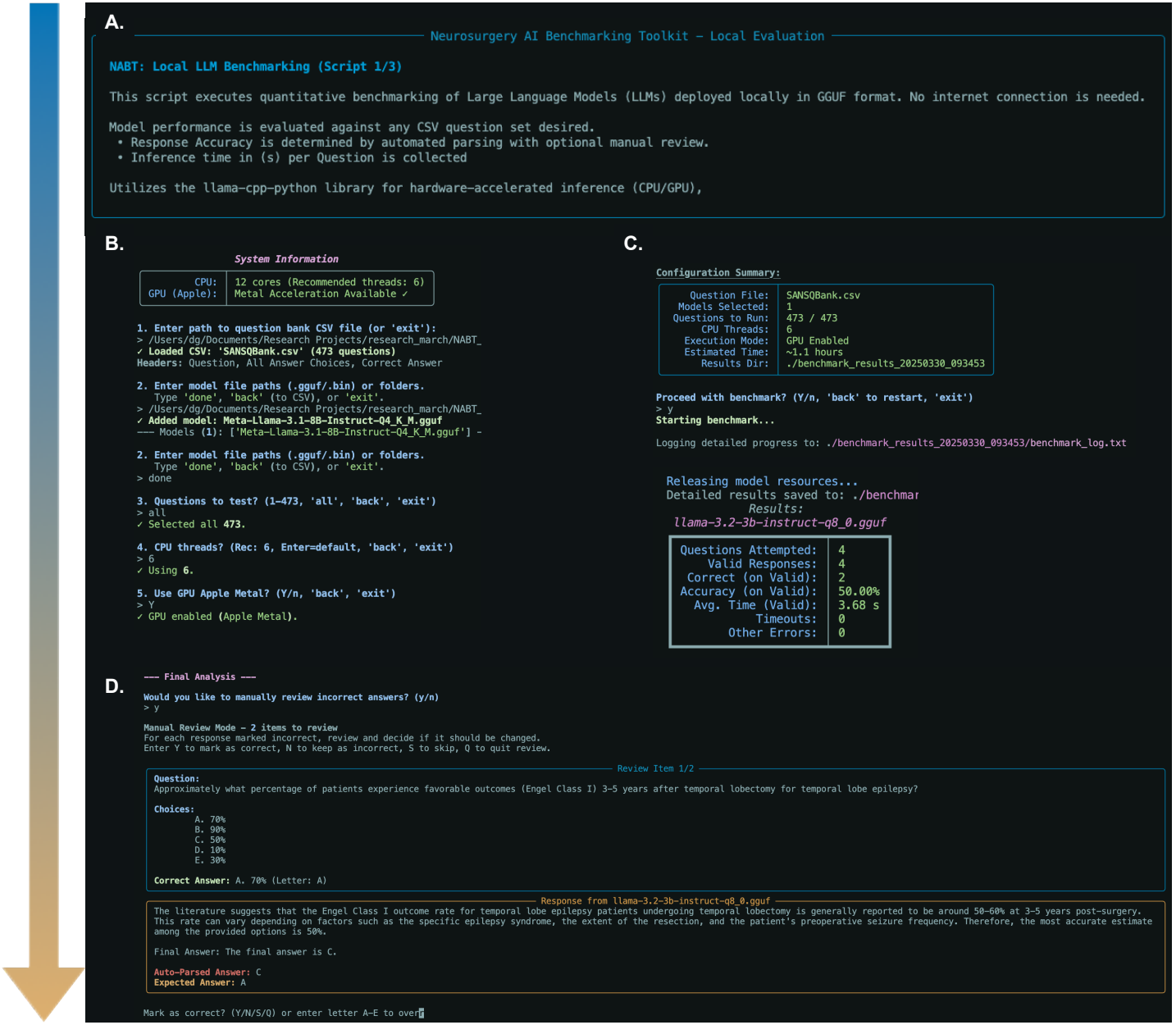

A.
Neurosurgery AI Benchmarking Toolkit – Local Evaluation
NABT: Local LLM Benchmarking (Script 1/3)
This script executes quantitative benchmarking of Large Language Models (LLMs) deployed locally in GGUF format. No internet connection is needed.
Model performance is evaluated against any CSV question set desired.
• Response Accuracy is determined by automated parsing with optional manual review.
• Inference time in (s) per Question is collected
Utilizes the llama-cpp-python library for hardware-accelerated inference (CPU/GPU),
B.
System Information
CPU: 12 cores (Recommended threads: 6)
GPU (Apple): Metal Acceleration Available ✓
1. Enter path to question bank CSV file (or 'exit'):
> /Users/dg/Documents/Research Projects/research_march/NABT_
✓ Loaded CSV: 'SANSQBank.csv' (473 questions)
Headers: Question, All Answer Choices, Correct Answer
2. Enter model file paths (.gguf/.bin) or folders.
Type 'done', 'back' (to CSV), or 'exit'.
> /Users/dg/Documents/Research Projects/research_march/NABT_
✓ Added model: Meta-Llama-3.1-8B-Instruct-Q4_K_M.gguf
--- Models (1): ['Meta-Llama-3.1-8B-Instruct-Q4_K_M.gguf'] -
2. Enter model file paths (.gguf/.bin) or folders.
Type 'done', 'back' (to CSV), or 'exit'.
> done
3. Questions to test? (1-473, 'all', 'back', 'exit')
> all
✓ Selected all 473.
4. CPU threads? (Rec: 6, Enter=default, 'back', 'exit')
> 6
✓ Using 6.
5. Use GPU Apple Metal? (Y/n, 'back', 'exit')
> Y
✓ GPU enabled (Apple Metal).
C.
Configuration Summary:
Question File: SANSQBank.csv
Models Selected: 1
Questions to Run: 473 / 473
CPU Threads: 6
Execution Mode: GPU Enabled
Estimated Time: ~1.1 hours
Results Dir: ./benchmark_results_20250330_093453
Proceed with benchmark? (Y/n, 'back' to restart, 'exit')
> y
Starting benchmark...
Logging detailed progress to: ./benchmark_results_20250330_093453/benchmark_log.txt
Releasing model resources...
Detailed results saved to: ./benchmar
Results:
llama-3.2-3b-instruct-q8_0.gguf
Questions Attempted: 4
Valid Responses: 4
Correct (on Valid): 2
Accuracy (on Valid): 50.00%
Avg. Time (Valid): 3.68 s
Timeouts: 0
Other Errors: 0
D.
--- Final Analysis ---
Would you like to manually review incorrect answers? (y/n)
> y
Manual Review Mode - 2 items to review
For each response marked incorrect, review and decide if it should be changed.
Enter Y to mark as correct, N to keep as incorrect, S to skip, Q to quit review.
Review Item 1/2
Question:
Approximately what percentage of patients experience favorable outcomes (Engel Class I) 3-5 years after temporal lobectomy for temporal lobe epilepsy?
Choices:
A. 70%
B. 90%
C. 50%
D. 10%
E. 30%
Correct Answer: A. 70% (Letter: A)
Response from llama-3.2-3b-instruct-q8_0.gguf
The literature suggests that the Engel Class I outcome rate for temporal lobe epilepsy patients undergoing temporal lobectomy is generally reported to be around 50-60% at 3-5 years post-surgery. This rate can vary depending on factors such as the specific epilepsy syndrome, the extent of the resection, and the patient's preoperative seizure frequency. Therefore, the most accurate estimate among the provided options is 50%.
Final Answer: The final answer is C.
Auto-Parsed Answer: C
Expected Answer: A
Mark as correct? (Y/N/S/Q) or enter letter A-E to over

A.

Demo Description

Demo 1: Private AI Retrospective/Prospective Database Helper
This demo takes one or more clinical notes and attempts to extract predefined data fields suitable for building a research database. The quality depends heavily on the note's structure and the model's ability.

You have selected **Meta-Llama-3.1-8B-Instruct-Q4_K_M.gguf** to demo.

Model Prompt Template

```
You are an AI assistant specializing in extracting structured clinical data from unstructured medical notes for a neurosurgical research database.
Analyze the provided clinical note(s) below. Extract the following specific pieces of information:

*   Patient MRN: (Medical Record Number, if available)
*   Date of Surgery: (YYYY-MM-DD format, if mentioned)
*   Procedure Name: (Specific name of the surgical procedure)
*   Tumor Size (mm): (Numerical value if mentioned, specify dimension if available e.g., APxWxH)
*   Tumor Location: (e.g., Left Frontal Lobe, Sellar, CP Angle)
*   EBL (cc): (Estimated Blood Loss in cc or mL)
*   Length of Stay (days): (If mentioned or calculable from dates)
*   Complications: (List any mentioned intra-operative or immediate post-operative complications, otherwise state 'None Reported')

Present the extracted data in a clear key-value list format below.
If a specific piece of information cannot be found in the provided text, clearly state 'Not Found'.

Clinical Note(s):
---
[Your pasted notes will go here]
---

Extracted Data:
```

You will provide your note or set of notes and evaluate the model's response.

⌘K to generate a command

B.

```
Generating response locally...
⁘ ⁘  Generating response...
```

C.

Interactive Demo: Clinical Note Extraction Performance

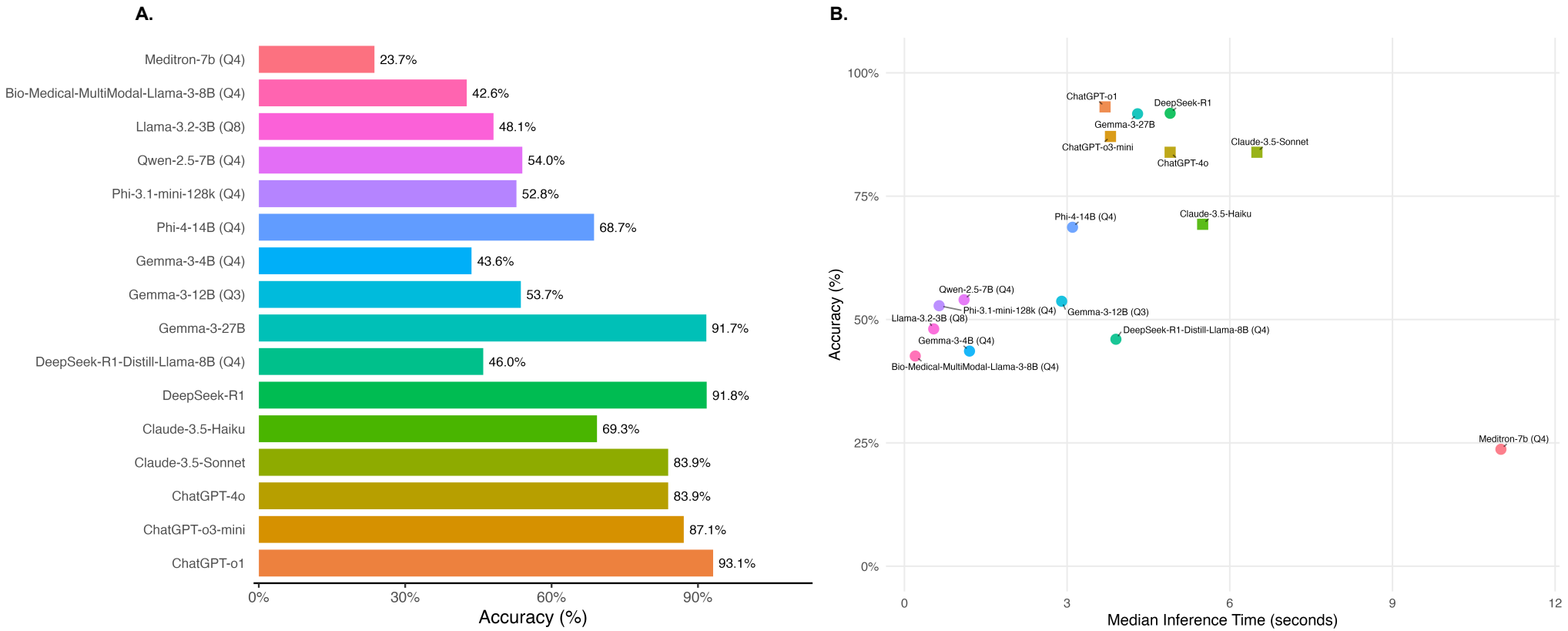
Subjective Quality Ratings (1-5 scale) by Three Independent Reviewers

Reviewer 1	Reviewer 2	Reviewer 3
Resident	*Medical Student*	*Medical Student*

Model	Rev 1	Rev 2	Rev 3	Avg
ChatGPT-o1 [API] SANS Accuracy: 93.1%	4.3	4.8	4.7	4.6
DeepSeek-R1 [API] SANS Accuracy: 91.8%	4.1	4.6	4.3	4.3
Phi-4-14B [Q4] SANS Accuracy: 68.7%	3.2	4.1	3.5	3.6
Qwen-2.5-7B [Q4] SANS Accuracy: 54.0%	2.7	3.4	2.1	2.7

Rating Scale

- 1-2 Poor
- 3 Average
- 4-5 Excellent


A.
Meditron-7b (Q4)
Bio-Medical-MultiModal-Llama-3-8B (Q4)
Llama-3.2-3B (Q8)
Qwen-2.5-7B (Q4)
Phi-3.1-mini-128k (Q4)
Phi-4-14B (Q4)
Gemma-3-4B (Q4)
Gemma-3-12B (Q3)
Gemma-3-27B
DeepSeek-R1-Distill-Llama-8B (Q4)
DeepSeek-R1
Claude-3.5-Haiku
Claude-3.5-Sonnet
ChatGPT-4o
ChatGPT-o3-mini
ChatGPT-o1
23.7%
42.6%
48.1%
54.0%
52.8%
68.7%
43.6%
53.7%
91.7%
46.0%
91.8%
69.3%
83.9%
83.9%
87.1%
93.1%
0%
30%
60%
90%
Accuracy (%)
B.
100%
75%
50%
25%
0%
Accuracy (%)
0
3
6
9
12
Median Inference Time (seconds)
ChatGPT-o1
DeepSeek-R1
Gemma-3-27B
ChatGPT-o3-mini
ChatGPT-4o
Claude-3.5-Sonnet
Phi-4-14B (Q4)
Claude-3.5-Haiku
Qwen-2.5-7B (Q4)
Phi-3.1-mini-128k (Q4)
Gemma-3-12B (Q3)
Llama-3.2-3B (Q8)
DeepSeek-R1-Distill-Llama-8B (Q4)
Gemma-3-4B (Q4)
Bio-Medical-MultiModal-Llama-3-8B (Q4)
Meditron-7b (Q4)