Using LLMs to score Hierarchical Organization of Narrative Recall

Hannah You, Jaylen Worthy, Ben Rickles[1], Donald Bolger[1,2]

1 Human Development, 2 Program in Neuroscience and Cognitive Science

PROGRAM IN NEUROSCIENCE & COGNITIVE SCIENCE

Introduction

Research has suggested that the ability to comprehend narratives relies on hierarchical organization of memory[1]. Disorders like dementia can be detected by measures of narrative recall structure; however these often rely on human graders [2, 3] . Here we explored how LLMs can be used to measure hierarchical organization of narrative recall (HONR). We looked at recall in healthy UMD subjects, who listened to stories in split into either scrambled or regular order. The scrambled stories attempted to simulate the conditions of dementia. We compared the grading measurements of LLMs and human scorers.

Hypotheses

1. Recall of Ordered stories > Recall of Scrambled stories
2. Can LLM **directly** measure narrative recall?
   - ChatGPT = Human Ratings
3. Can LLM +coding Indirectly measure narrative recall?
   - Hierarchical (EventBunching) = Global Coherence
   - Linear (Linearity) = Local Coherence

Methods & Materials

- ❑ Subjects: 5 UMD students age 18-22
- ❑ Procedure:
  - ❑ 2 Narrative Blocks, 2 Conditions
    - ❑ Ordered Story (Sentences in sequential order)
    - ❑ Scrambled Story (Sentences Out-of-order)
  - ❑ Each 45-55 Sentences
  - ❑ Each 10-11 Events
  - ❑ ~5 minutes each
- ❑ Verbal Recall (As many details as possible)

Scoring procedure

- ❑ Transcribe recall
- ❑ **Human ratings** [2,3]
  - **Local Coherence:** (Connective phrases)
  - **Global Coherence:** (Existence of Themes)
- ❑ Direct Measures
  - Via Query to ChatGPT
- ❑ **Indirect LLM ratings** (Google Colabs ipynb)
  - Derive **Referenced Recall Clauses**
    - Via Query to ChaptGPT
  - **Linearity:** How much does the recall proceed in a linear fashion relative to the story?
    - (Total clauses after each clause in the recall which chatGPT attributed to sentences later in the story)
  - **Event Clustering:** How often do clauses of recall cluster around single events?
    - (number of occasions in which event IDs of adjacent recall clauses attributed by chatGPT match/mismatch)

1 Human Ratings

Local Coherence: each sentence was graded on a scale of 0-1 based on whether it connected to the previous sentence. This was determined on four measures: sequencing adverbials, pronominal reference, definite determiners, and statements of cause and effect.[2,3]

Global Coherence: the stories were graded on a 1-4 scale based on whether they satisfied the 4 core themes of the stories.[2,3]

Sentences	Local Coh	Global Coh
Um, Let's see. So there was a prince who walked into into a cave.	NA	1
The cave was holy I think, and like people thought it was.	1	0
And so as he walked through the cave he looked around and saw..	1	1
And there was something about a pool, I don't remember.	0	0
There was something about a traveller and a dog.	0	0

Direct LLM Recall Measures

us:
1. I have this recall of a story …..
2. Can you rate the local coherence of the recall on a 0-1 scale?
3. Can you rate the prominence of the on following theme within the recall on a 1-4 scale …

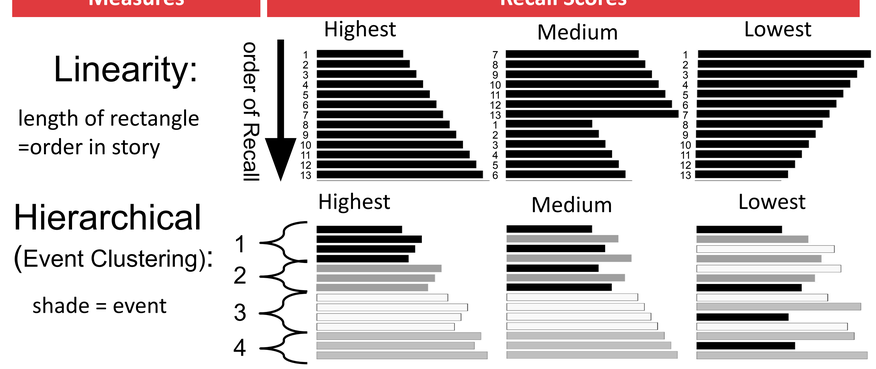
GPT:
1. Sure, here's X and Y:...

Indirect LLM Scoring (LLM + Code)

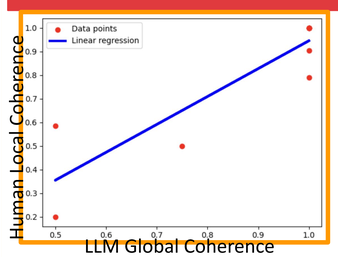
us:
1. I have a story with the following sentences in the following order:..[Story]
2. I have this recall of the story:..... [Recall]
3. Can you please label each of the clauses in the recall sentences with their corresponding sentence in the original story in the order that they appear in the recall?

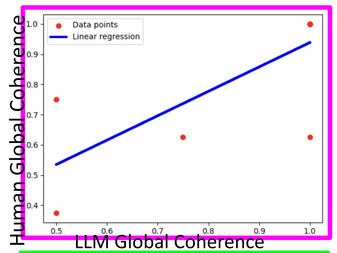
GPT:
1. Sure, here's a breakdown of each recall sentence with its corresponding sentence from the original story:...

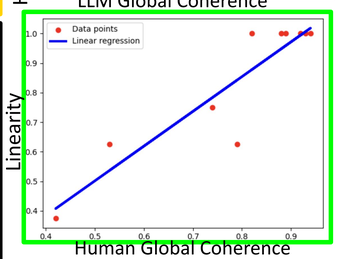
Python: Story Sentences + Referenced Recall Clauses + Code = Scores



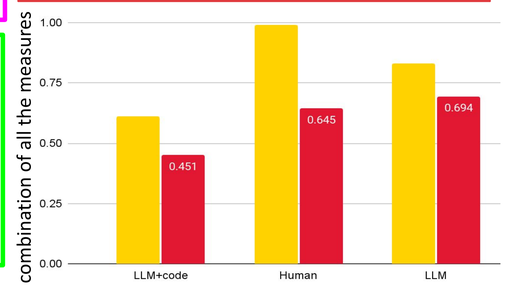
Results: Human vs. LLM scores







Results: Scrambled vs Regular



Summary

1. Recall of Ordered stories was > Recall of Scrambled on all measures
2. Direct ChatGPT ~correlated with Human ratings
   - Global – Yes
   - Local – No
3. Indirect (LLM + Coding) ChatGPT ~correlated with Human Ratings
   - Linearity= Correlated w/ Human Global & Local Coherence
   - Hierarchical not correlated

This promises multiple ways LLM can score hierarchical organization of narrative recall with an accuracy comparable to human diagnostics.

Limitations / Future Directions

- LLMs adapt and are different. Grading inconsistent across LLMs/time
- Abstract definition and operalization of Global coherence
- Some LLM error:
  - LLMs can sometimes be fooled by fabricated recall sentences.
  - Connecting story sentences to recall clauses: 16.12% error
- Low sample size!
- How to validate this for diagnostic purposes?
  - different populations
  - story lengths

References

[1] Farag, C., Troiani, V., Bonner, M., Powers, C., Avants, B., Gee, J., & Grossman, M. (2010). Hierarchical Organization of Scripts: Converging Evidence from fMRI and Frontotemporal Degeneration. Cerebral Cortex, 20(10), 2453–2463. https://doi.org/10.1093/cercor/bhp313

[2] Ash, S., Moore, P., Antani, S., McCawley, G., Work, M., & Grossman, M. (2006). Trying to tell a tale Discourse impairments in progressive aphasia and frontotemporal dementia. In NEUROLOGY (Vol. 66). www.neurology.org

[3] Ash, S., Xie, S. X., Gross, R. G., Dreyfuss, M., Boller, A., Camp, E., Morgan, B., O'Shea, J., & Grossman, M. (2012). The organization and anatomy of narrative comprehension and expression in Lewy body spectrum disorders. Neuropsychology, 26(3), 368.