Memory for Long Narratives

M. Toneva[1,2], V. Vo[3], J. Turek[3], S. Jain[4], S. Michelmann[1], M. Capotă[3], A. Huth[4], U. Hasson[1], and K. A. Norman[1]

[1]Princeton University, Princeton, USA [2]Max Planck Institute for Software Systems, Saarbrücken, Germany
[3]Intel Labs, Hillsboro, USA [4]University of Texas at Austin, Austin, USA

Summary

Background: Language is the primary way in which we communicate, and yet it is not clear how we draw on previous experiences and integrate information over long timescales to understand language.

Goal: Investigate the role of episodic memory in language comprehension, by building models of this process and by collecting new benchmark datasets.

This work: Progress towards a large dataset of memory performance for long narratives, and a scalable, automated scoring of memory performances.

Findings:

A number of events from an intermediate chapter of a 300 page novel were recalled with high precision across participants

Automated scoring of recall can be enabled by recent natural language models (e.g. GPT-2)

Motivation

Much work in language comprehension focuses on the word- and sentence-level

Something richer happens at the narrative-level [1,2]

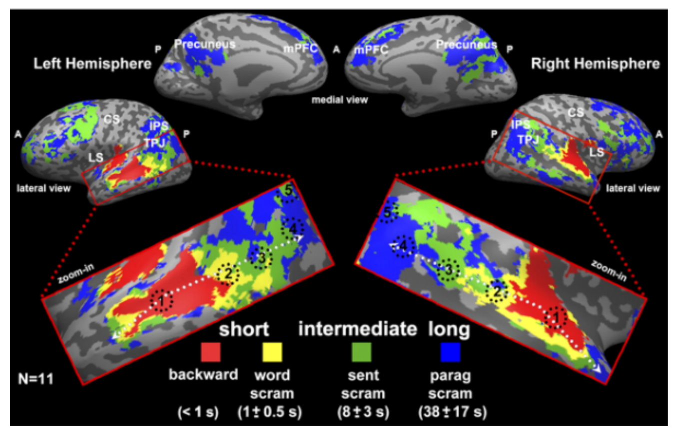
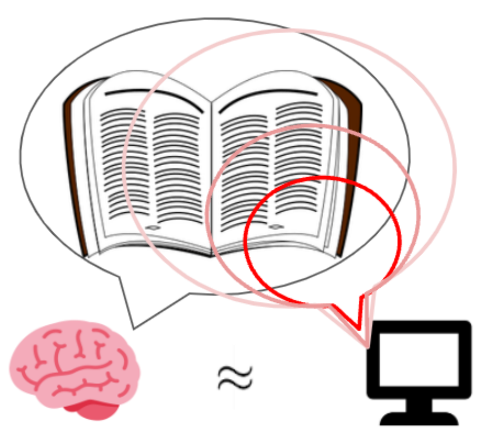


Fig. from [1]



Long-term goals

Study the role of episodic memory in understanding **long narratives**

Contrast **human vs. SOTA natural language processing** (NLP) model memory performance

This work makes progress towards the following requirements for this direction:

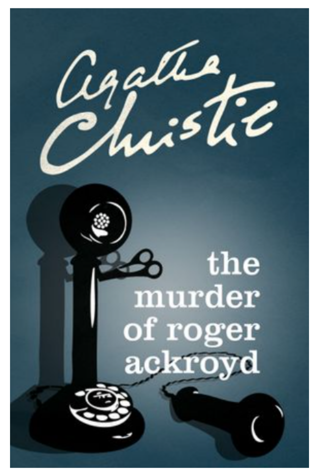
Large dataset of human memory performance for **long narratives** (i.e. books)

Automated scoring of memory performance to enable scaling to millions of datapoints

Acknowledgments

This work was made possible by support from the Ruth L. Kirschstein National Research Service Award Institutional Research Training Grant T32MH065214 and Princeton University's C.V. Starr Fellowship.

Behavioral Dataset



Task: recall chapters of a recently-read novel when cued with a passage from the start of the chapter

Novel: *The Murder of Roger Ackroyd* by Agatha Christie [3]

288 pages
27 chapters
Words per chapter: 2590.1 ± 172.6 words

Recruitment: online forums, within 2 mo. of finishing

Data collection is ongoing. Preliminary results based on n=15 for Chapter 22.

Dataset Statistics

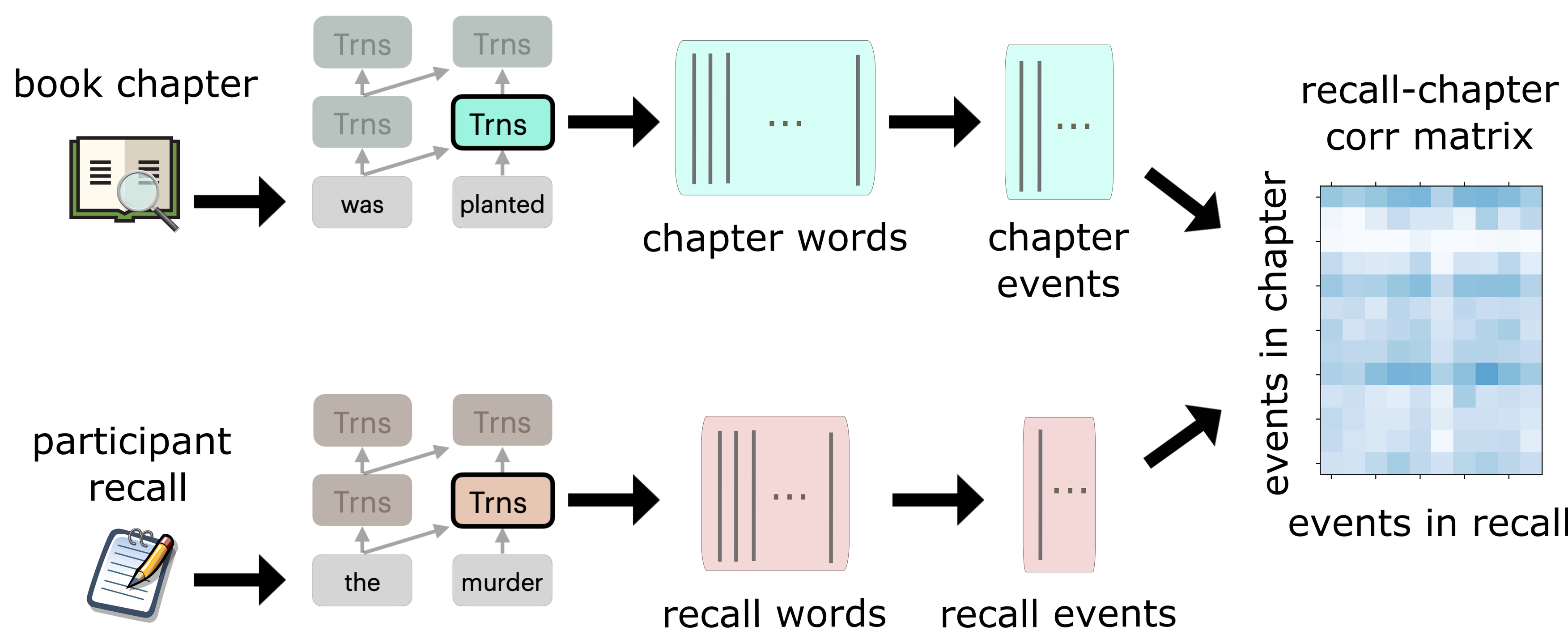
N=15 (ongoing)
Recall length: 55.7 ± 8.8 words
Days since finished: 28.4 ± 6.5 days
Modality: 13 Read, 2 Listened

Demographics

Reading frequency: 10+ books/yr
Age: 33.2 ± 2.8 years
Sex: 15 F
Language: 13 Native E, 2 Fluent E

Recall Analysis

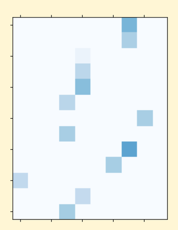
Automated scoring of recall based on word representations extracted from a large neural network NLP model (GPT-2 [4]), inspired by Heusser et al. 2021 [5]



Goal: Characterize how **different decision points** affect how accurately the automatic scoring estimates well-recalled events

Scoring metrics?

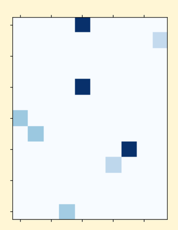
Precision [5]



best recall match:

max of each row

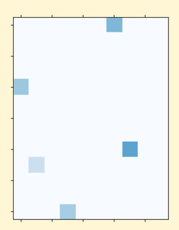
Distinctiveness [5]



penalize recall events with multiple matches:

z-score within col. then max of each row

Corr. of best unique matches



greedy matching:

find global max, zero out the rest of its row & col., then find next largest value, repeat

What model, layer, context length?

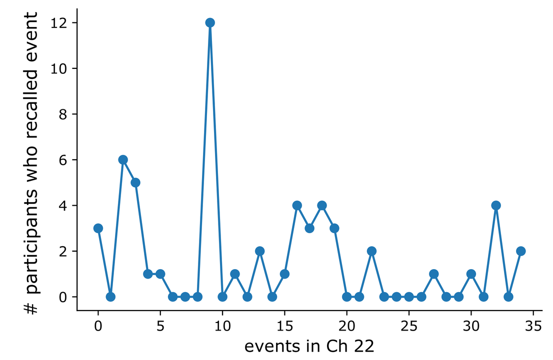
How to aggregate within event?

Baseline: hand scoring

1. Annotate event boundaries in chapter text and recall
2. For each recall event, assign most semantically-related chapter event

Results

Baseline: hand **annotations**

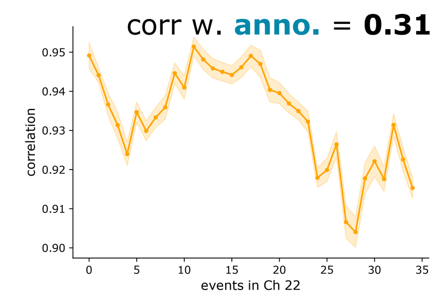


8 of 35 chapter events recalled in at least 20% of participants

One event recalled in 80% of participants related to an important plot twist

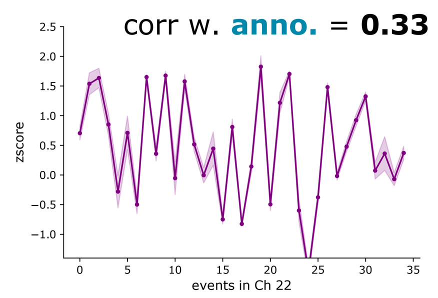
Partially-automated scoring (events segmented by hand)

Precision



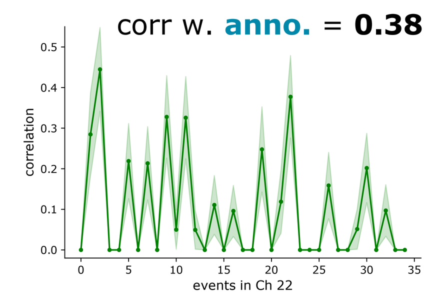
model: GPT-2
layer: 11 of 12 hidden
aggregation: mean pool

Distinctiveness



model: GPT-2
layer: 10 of 12 hidden
aggregation: last word

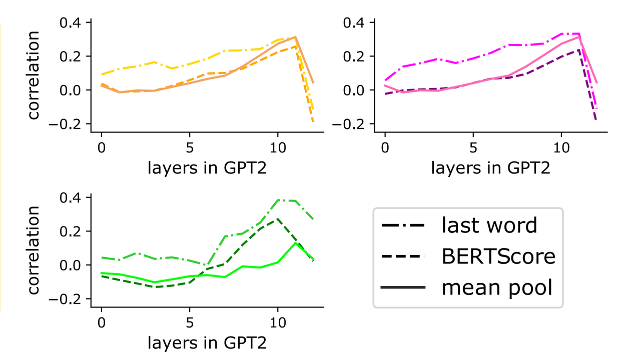
Corr. of best unique

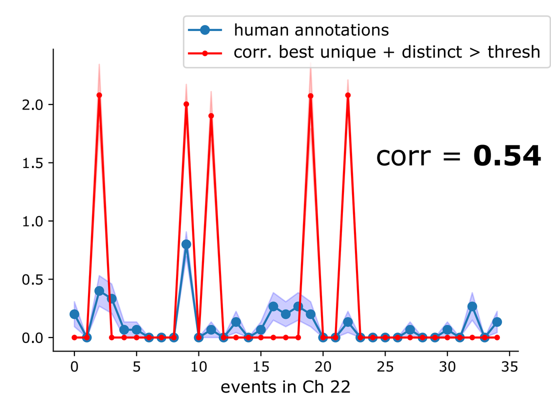


model: GPT-2
layer: 10 of 12 hidden
aggregation: last word

Metrics are **affected by layer depth & aggregation method**

- deeper layers >> shallow layers
- last word aggregation is best





Combining two metrics and applying a threshold can improve automated scoring

- higher corr. with annotations
- no false positives

Conclusion

Scalable automated scoring of recall is enabled by recent NLP models, but its quality is **affected by various factors**:

layer of extraction
method of aggregation within semantically meaningful units
metric for comparison with original chapter text

Next steps are to fully automate scoring by relaxing the dependence on event boundaries segmented by humans, and test on a wider set of chapters and on NLP model-generated recall.

References

[1] Lerner, Yulia, Christopher J. Honey, Lauren J. Silbert, and Uri Hasson. "Topographic mapping of a hierarchy of temporal receptive windows using a narrated story." Journal of Neuroscience 31, no. 8 (2011).
[2] Wehbe, Leila, Brian Murphy, Partha Talukdar, Alona Fyshe, Aaditya Ramdas, and Tom Mitchell. "Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses." PloS one 9, no. 11 (2014): e112575.
[3] Christie, Agatha. The Murder of Roger Ackroyd. 1926.
[4] Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. "Language models are unsupervised multitask learners." OpenAI blog 1, no. 8 (2019): 9
[5] Heusser, Andrew C., Paxton C. Fitzpatrick, and Jeremy R. Manning. "Geometric models reveal behavioural and neural signatures of transforming experiences into memories." Nature Human Behaviour 5, no. 7 (2021).