Automated recall-to-story matching for memory research

Dhruva Arekar[1], Gabriel Kressin Palacios[1]*, Xian Li[1], Kahlyn Eckles[2], Viswanath Missula[1], Janice Chen[1]*

[1]Department of Psychological & Brain Sciences, Johns Hopkins University
[2]Department of Psychology, University of Oregon
** Equal contribution*

Introduction

> Scientists are increasingly interested in using free recall of stories to study human memory[1]

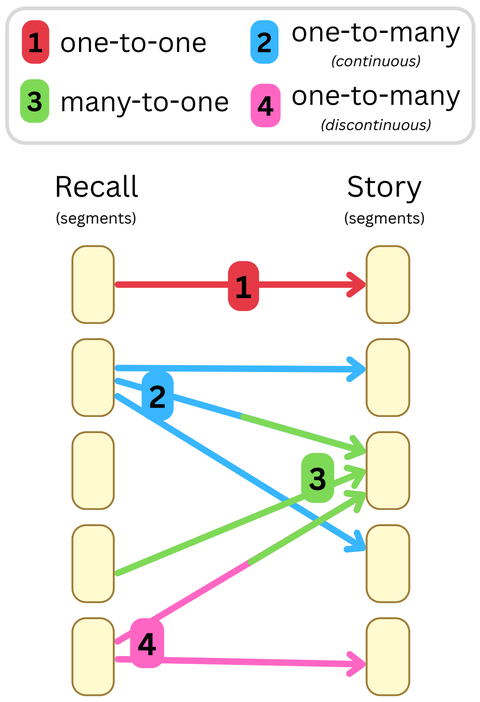
> Scoring a story recall test requires matching sections of the recall transcript to the story elements being described

> Matching is usually performed by humans (laborious, susceptible to error, sensitive to instructions)[2-5]

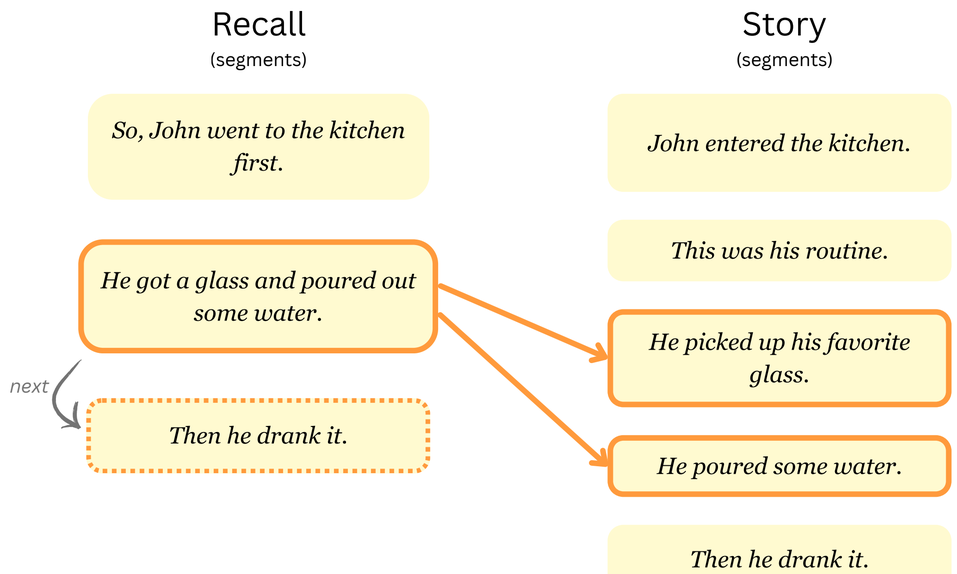
> Prior work has partly automated matching, but does not provide a validated, general, and standalone tool[6-12]

We aim to automate and validate a pipeline for recall-to-story matching.

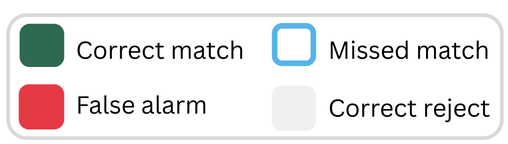
Types of Matches

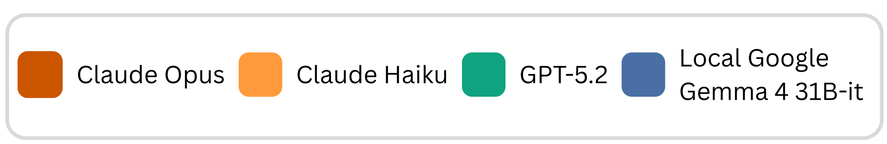


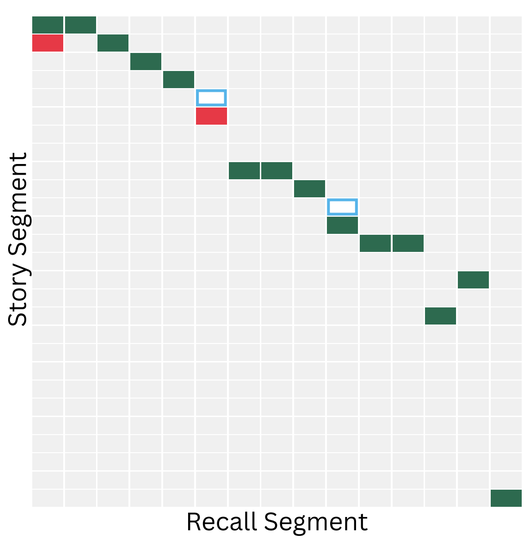
Matching Paradigm

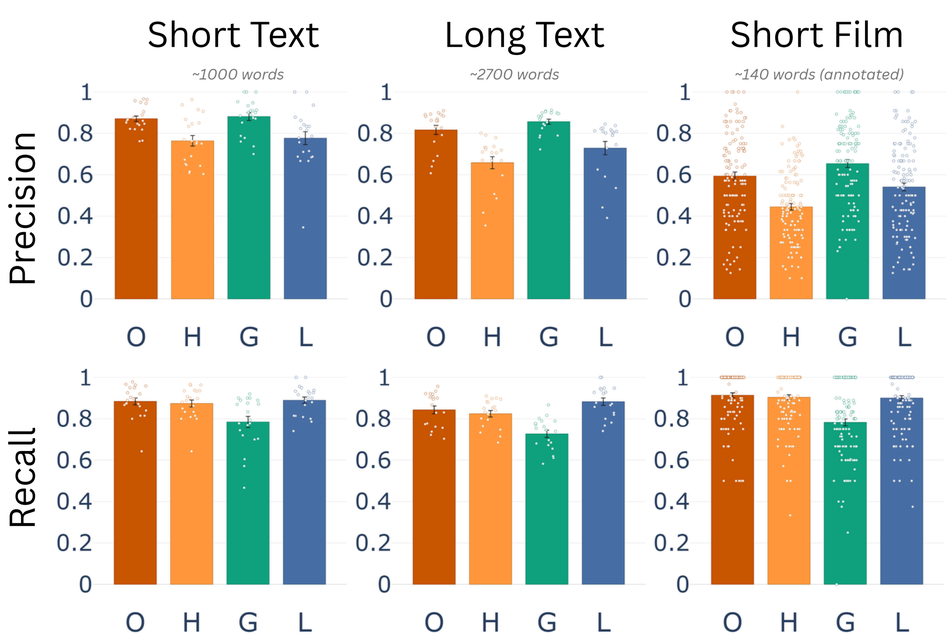


`rMatch` performs similarly to humans

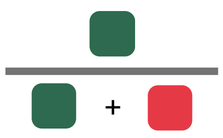




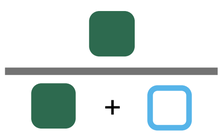




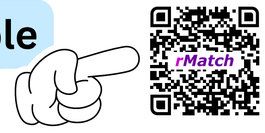
Precision



Recall



Find `rMatch` as an importable Python module online!



Ongoing scaling of evaluation

> To what degree do humans agree on matches?

> We are building an open benchmark dataset to assess inter-rater reliability

> We investigate across multiple modalities and story lengths

	Read	Listen	Film
Short	**45 recalls** 3 stories *222 w*	**~ 75 recalls** 15 stories *200 w*	**30 recalls** 10 stories *1-2 min*
Medium	**6 recalls** 2 stories *1000 w*	**27 recalls** 8 stories *600-1500 w*	**21 recalls** 7 stories *2-7 min*
Long	**6 recalls** 2 stories *2700 w*	**3 recalls** 1 story	**6 recalls** 2 stories *50-108 min*

Conclusions

> We introduce rMatch, a Python-importable and evaluated package for automated matching

> rMatch assigns matches similarly to humans
- > Human-assigned matches are rarely missed (high *recall*)
- > Most mistakes are overeager matches (lower *precision*)

> Fully local performance almost on par with high-end Claude models (works with 94GB of VRAM)

Future Directions

> Calculate human inter-rater reliability (the expected ceiling for rMatch)

> Quantify *how well* a recall segment conveys a matched story segmented (e.g. with mutual information)

> Extend rMatch to work directly on movies using video encoders

References

1. Lee et al. *Current Opinion in Behavioral Sciences* (2020)
2. Baldassano et al. *Neuron* (2024)
3. Antony et al. *Journal of Cognitive Neuroscience* (2024)
4. Barnett et al. *Neuron* (2024)
5. Lee & Chen *Nature Communications* (2022)
6. Panela et al. *Communications Psychology* (2025)
7. Mu et al. b*ioRxiv* (2025)
8. Chen, Raccah et al. *PsyArXiv* (2025)
9. Georgiou et al. *Learning & Memory* (2025)
10. Zhong et al. *Physical Review Letters* (2025)
11. Heusser et al. *Nature Human Behaviour* (2021)
12. Lee et al. *PsyArXiv* (2024)