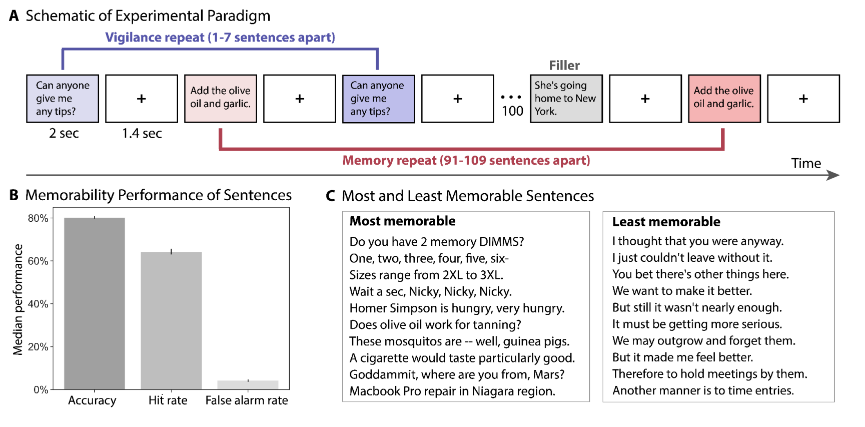
# Meaning distinctiveness as a key predictor of sentence memorability within a noisy representation account

Thomas Hikaru Clark, Greta Tuckute, Bryan Medina, Evelina Fedorenko
Department of Brain and Cognitive Sciences, MIT.

**Background.** Prior work on visual memory suggests that memory performance is best explained by memory representations accumulating random noise over time, and the mind employing probabilistic judgments of familiarity as opposed to an exact, all-or-none lookup [1]. Here, we test this *noisy representation hypothesis* on linguistic stimuli to see if it explains variation in sentence memorability. This hypothesis predicts that sentences with more distinctive semantic representations are less prone to false alarms and false negatives, and should be faster for participants to recognize, due to reduced confusion with similar distractors. We test these predictions while operationalizing semantic distinctiveness for sentences and controlling for word-level features, given that single words and compositional phrases may be remembered differently [2]. **Methods.** Drawing on past work on memorability for faces, images, and words [3,4,5], we conducted a recognition memory experiment (*Fig.* 1) with N=500 native English speakers on Prolific. Materials consisted of 2500 six-word target sentences from diverse English-language corpora, split into 3 groups of 500 based on word-level properties and 1000 broadly sampled "diverse" sentences. As word-level controls, we compute estimated word memorability based on the empirical results of a previous study [5] ("word memorability") and mean word frequency. To quantify semantic distinctiveness at the sentence level, we map sentences to a semantic vector space using the Sentence-BERT large language model [6], and compute each sentence's average cosine distance from other sentences in the source corpora. We use linear regression to predict sentences' mean response accuracy (i.e. the number of hits and correct rejections out of all presentations of a sentence across participants), and linear mixed-effects regression to predict reaction times. We also investigate the influence of distinctiveness (relative to potential distractors) on false positives, using logistic mixed-effects regression to predict correct rejections on non-repeat trials. **Results.** Sentences show moderate inter-participant correlations in accuracy (median split-half correlation=0.56). Average word frequency (b=-0.101, SE=0.034, p=.003), word memorability (b=0.328, SE=0.042, p<.001), and semantic distinctiveness (b=0.314, SE=0.029, p<.001) are all significant predictors of accuracy (*Fig.* 2). Word memorability and semantic distinctiveness showed a significant negative interaction (b=-0.158, SE=0.034, p<.001), suggesting that distinctiveness has a larger effect when a sentence's constituent words are not memorable (but such words may combine to form distinctive meanings). Additionally, both word memorability (b=-40, SE=3.5, p<.001) and semantic distinctiveness (b=-32, SE=2.9, p<.001) had negative effects on reaction time. The maximum cosine similarity of previously appearing sentences (b=-0.59, SE=0.05, p<.001) and the presence of a previously appearing content word (b=-0.45, SE=0.06, p<.001) had negative effects on correctness by increasing false positives, which suggests that distinctiveness relative to recently processed items — not just relative to sentences in general — boosts memorability (*Fig.* 3). **Conclusion.** We show that distinctiveness serves as a key predictor of sentence memorability, and that it can be operationalized using the semantic representation space of large language models. Our results are consistent with the noisy representation hypothesis, whereby recognition memory performance (both accuracy and speed) depends on the level of uncertainty surrounding an item's familiarity; for a sentence, this can be modeled by how distinctive its overall meaning is both from other sentences in the experiment and sentences more broadly.



**Most memorable**

Do you have 2 memory DIMMS?
One, two, three, four, five, six-
Sizes range from 2XL to 3XL.
Wait a sec, Nicky, Nicky, Nicky.
Homer Simpson is hungry, very hungry.
Does olive oil work for tanning?
These mosquitos are -- well, guinea pigs.
A cigarette would taste particularly good.
Goddammit, where are you from, Mars?
Macbook Pro repair in Niagara region.

**Least memorable**

I thought that you were anyway.
I just couldn't leave without it.
You bet there's other things here.
We want to make it better.
But still it wasn't nearly enough.
It must be getting more serious.
We may outgrow and forget them.
But it made me feel better.
Therefore to hold meetings by them.
Another manner is to time entries.

**Fig. 1 (A)** Experimental paradigm consisting of serially presented sentences and a repeat detection task. **(B)** Accuracy, hit rate, and false alarm rate across experimental materials. **(C)** The most and least memorable sentences among the experimental materials.

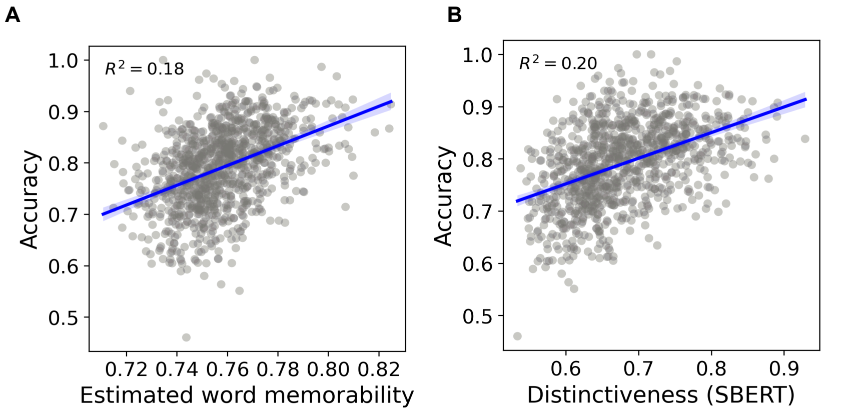


**Fig. 2.** Each point denotes an experimental sentence within the n=1000 “diverse” set. Both estimated word memorability **(A)** and SBERT distinctiveness **(B)** are correlated with mean response accuracy (for an item, the proportion of hits and correct rejections out of all presentations across participants). Both have significant effects on accuracy when included as covariates in a linear regression, with a significant negative interaction.

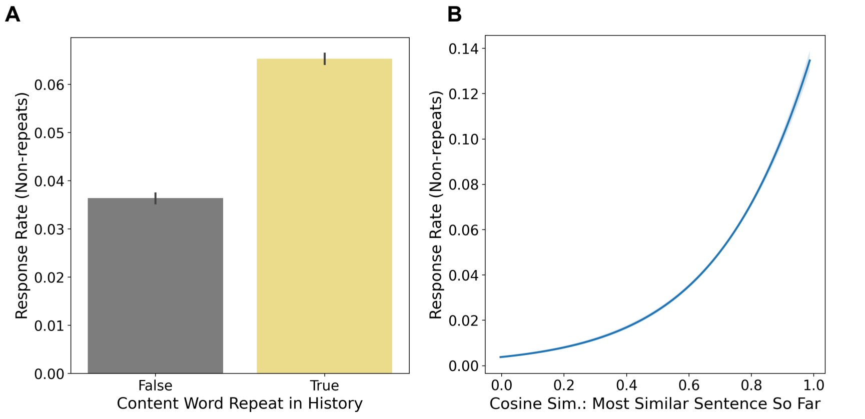


**Fig. 3. (A)** Comparison of response rate on non-repeats for trials with and without a content word repeat in the participant-specific experimental sequence. Error bars denote 95% CIs. **(B)** Logistic fit of responses on non-repeat trials as a function of the SBERT cosine similarity between the current sentence and the most similar sentence appearing so far. Shaded region (very narrow) denotes 95% confidence interval.

**References:**

**[1]** Brady, T. F., Robinson, M. M., & Williams, J. R. (2024). Noisy and hierarchical visual memory across timescales. Nature Reviews Psychology. **[2]** Jacobs, C. L., Dell, G. S., Benjamin, A. S., & Bannard, C. (2016). Part and whole linguistic experience affect recognition memory for multiword sequences. JML. **[3]** Bainbridge, W. A. (2017). The memorability of people: Intrinsic memorability across transformations of a person’s face. JEP. **[4]** Isola, P., Xiao, J., Torralba, A., Oliva, A. (2011) What makes an image memorable? CVPR. **[5]** Tuckute, G., Mahowald, K., Isola, P., Oliva, A., Gibson, E., & Fedorenko, E. (2024). Intrinsically memorable words have unique associations with their meanings. Preprint. **[6]** Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. arXiv.