# So you want to model cognition?

A Guided Tour with MINERVA2

**Sydelle de Souza**

CCS Guest Lecture | 10th November 2025

THE UNIVERSITY *of* EDINBURGH
UKRI Centre for Doctoral Training in Natural Language Processing

UK Research and Innovation

THE UNIVERSITY *of* EDINBURGH
School of Philosophy, Psychology and Language Sciences

# #1

# Every modeling choice is a theoretical commitment

# #2

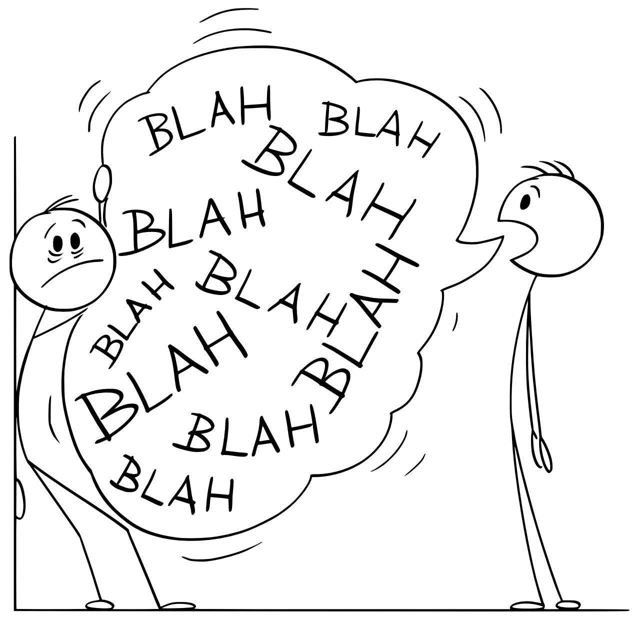
# Cognitive modeling starts with a research question, not a dataset.

# BACKGROUND & RQS

Human language is fundamentally compositional.

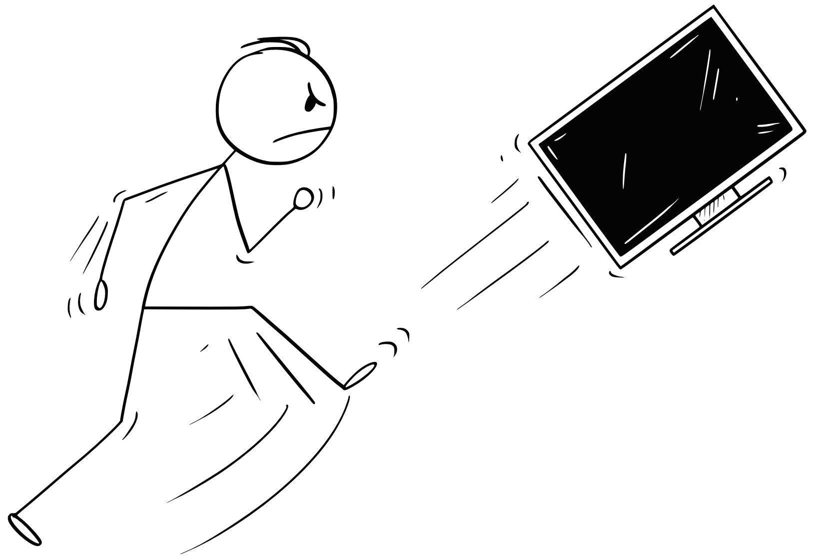
Chomsky, N. (1965). Aspects of the Theory of Syntax. MIT Press.

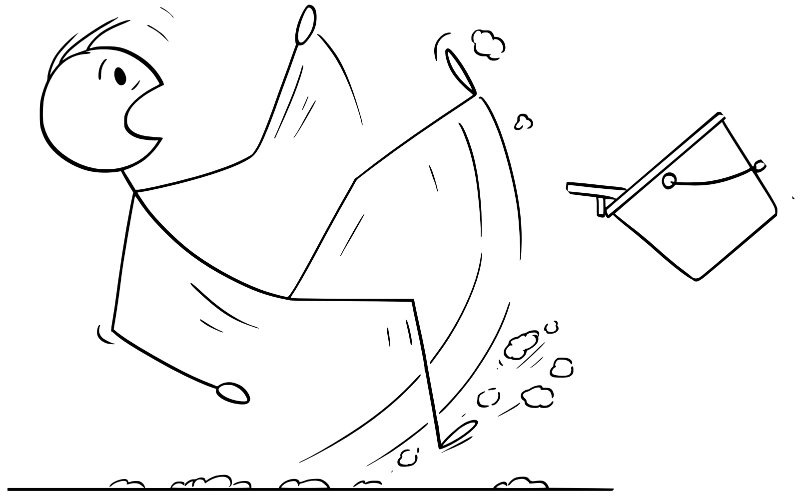
**LINGUISTIC UNITS** + **LINGUISTIC RULES** →

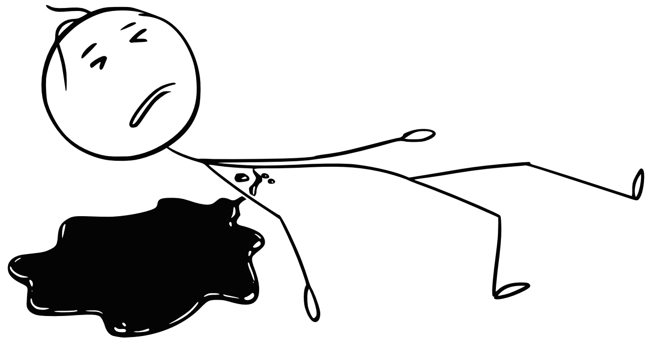


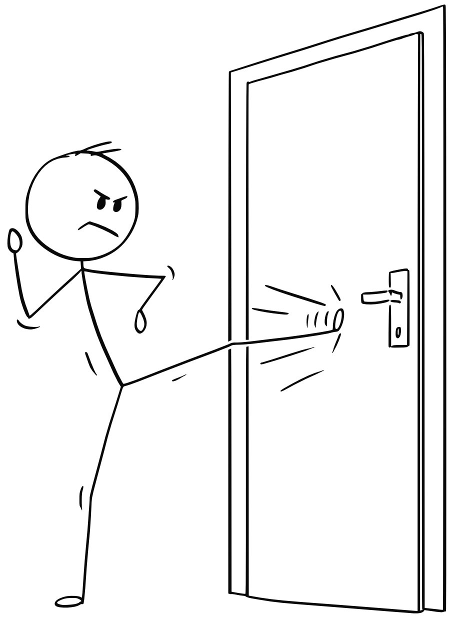

*infinite, novel utterances*

Chomsky, N. (1965). Aspects of the Theory of Syntax. MIT Press.



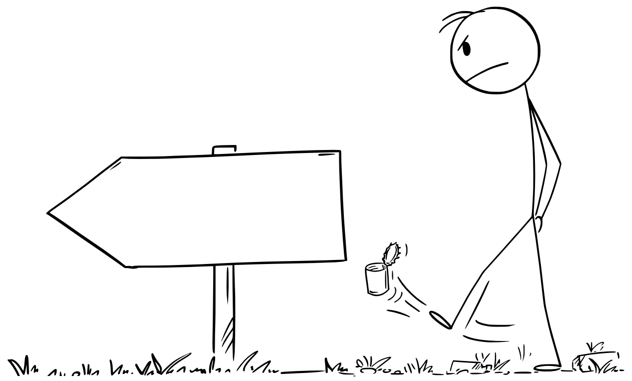


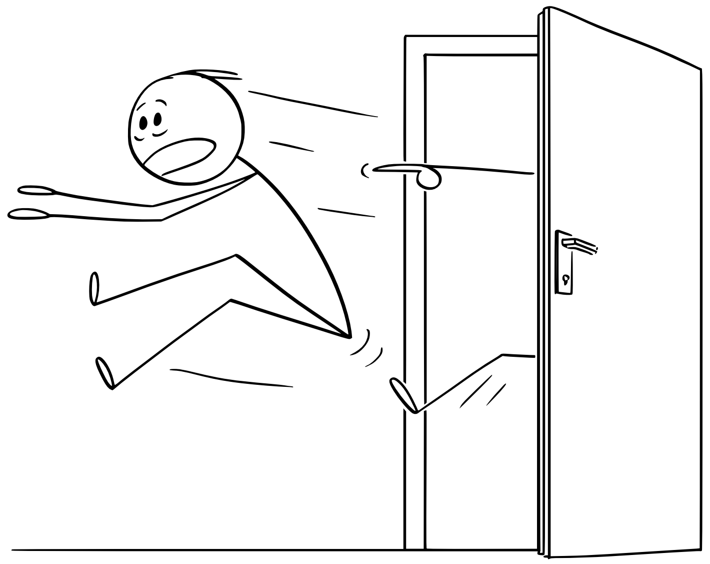


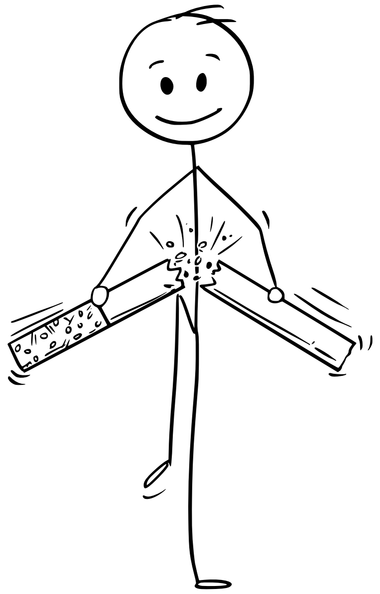


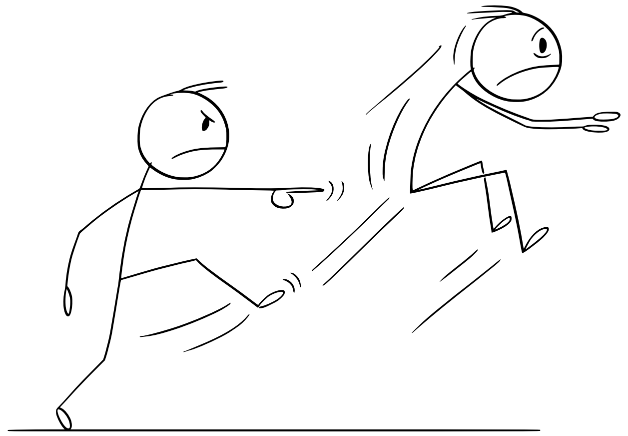
**to kick**

(v. transitive)









# SEMANTIC COMPOSITIONALITY

The degree to which the meanings of the constituent words contribute to the overall meaning of the expression.

+ COMPOSITIONAL

- COMPOSITIONAL

**PRODUCTIVE COMBINATIONS**

**kick + ball**

Literal Verb

Literal Noun

**COLLOCATIONS**

**kick + habit**

Figurative Verb

Literal Noun

**IDIOMS**

**kick + bucket**

Figurative Verb

Figurative Noun

Howarth, P. (1998). Phraseology and second language proficiency. Applied Linguistics, 19(1), 24–44.
https://doi.org/10.1093/applin/19.1.24

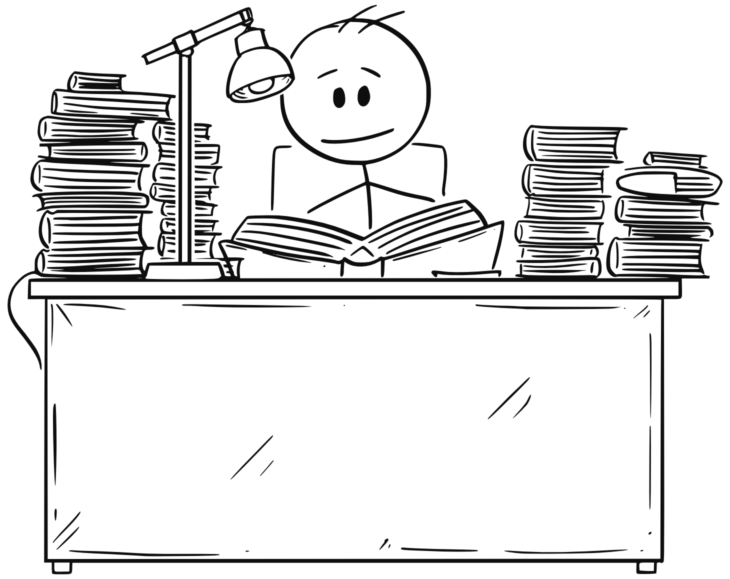
COLLOCATIONS

# kick + habit

Figurative
Verb

Literal
Noun

How do humans **acquire** and **process** semi-compositional language?



**Exposure**

freeze {🐟, 💧, 🍛} → {🐟, 💧, 🍛}
becomes cold enough to become solid

↓

**Abstraction**

freeze (v.)
to become cold enough to become solid
[+ reversible]

↓

**Generalization**

freeze {🥩, 🥛 } → {🥩, 🥛}
becomes cold enough to become solid

**Exposure**

↓

**Abstraction**

↓

**Generalization**

freeze {🐟, 💧, 🍛} → {🐟, 💧, 🍛}

becomes cold enough to become solid

freeze (v.)

to become cold enough to become solid

**liquid → solid**

[+ reversible]

freeze {🥩, 🥛, N (7, Nitrogen, 14.00674), 🌋} → {🥩, 🥛, N (7, Nitrogen, 14.00674), 🌋}

becomes cold enough to become solid

**Exposure**

freeze {🐟, 💧, 🍛} → {🐟, 💧, 🍛}

becomes cold enough to become solid

↓

**Abstraction**

freeze (v.)

to become cold enough to become solid

liquid → solid

**fluid → rigid**

[+ reversible]

↓

**Generalization**

**freeze {account, time} → {account, time}**

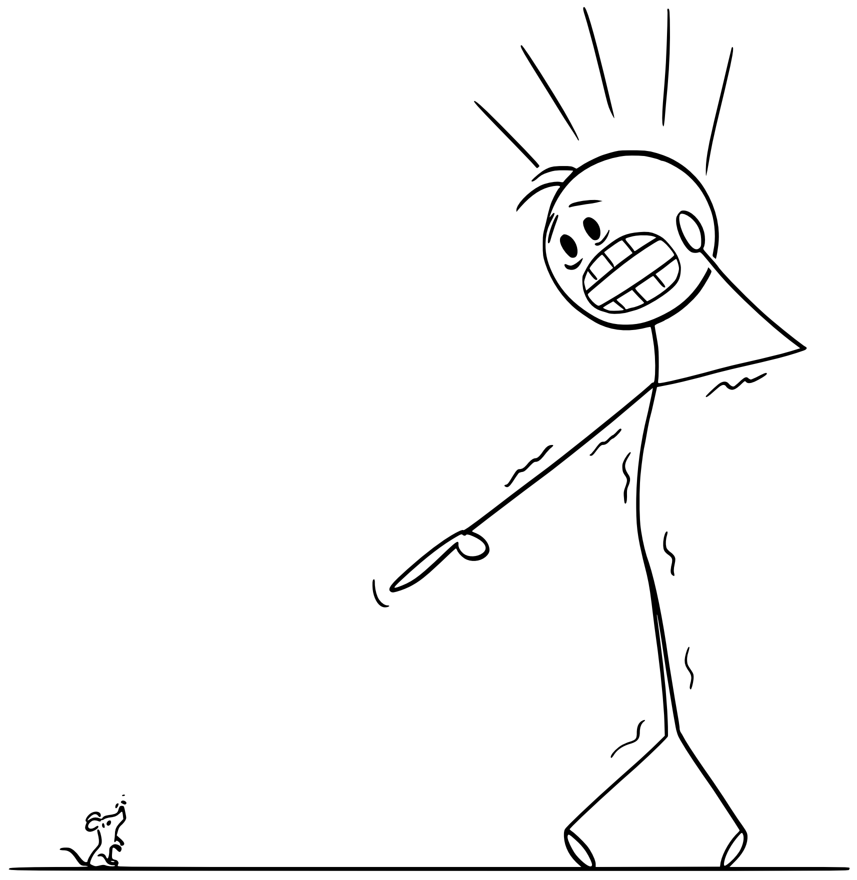
# LEARNING

**freeze** LITERAL

- to become cold enough to become solid
- liquid → solid
- **fluid → rigid**
- **not permanent; reversible**

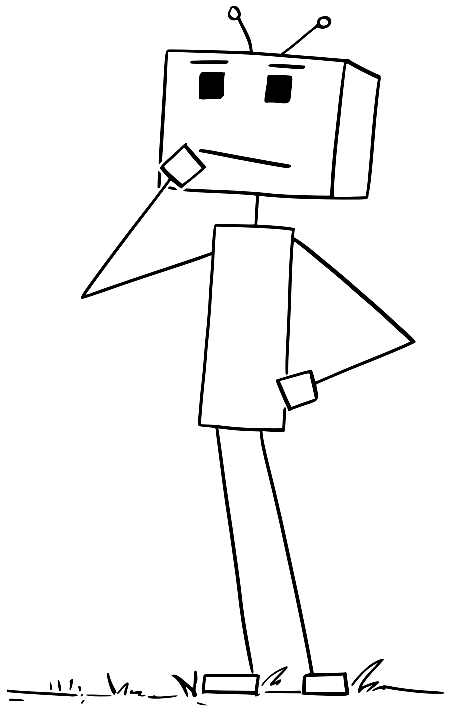
**freeze** FIGURATIVE

**to temporarily immobilise {x}**



# COLLOCATIONS POSE A LEARNING CHALLENGE

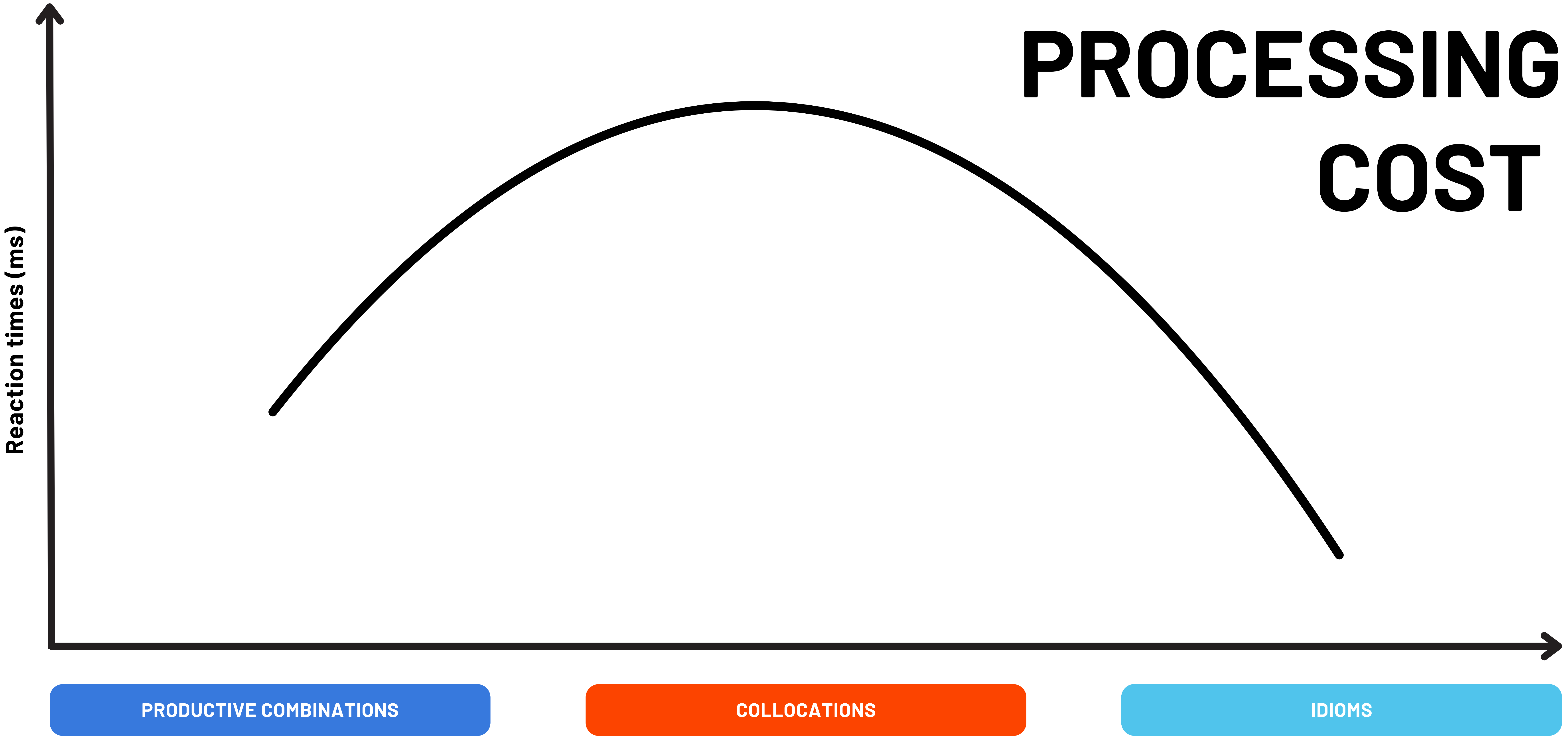
**L2 speakers**



**Machines**

Mi, M., Villavicencio, A., & Moosavi, N. S. (2025). Rolling the DICE on Idiomaticity: How LLMs Fail to Grasp Context. In W. Che, J. Nabende, E. Shutova, & M. T. Pilehvar (Eds), Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 7314–7332). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.acl-long.362
Howarth, P. (1998). Phraseology and second language proficiency. Applied Linguistics, 19(1), 24–44. https://doi.org/10.1093/applin/19.1.24

# PROCESSING COST




Imaginary graph. Not to scale.

# FREQUENCY

- Roughly half of human language is comprised of compositional units
- Collocations are the largest subset of figurative language

+ COMPOSITIONAL

- COMPOSITIONAL

PRODUCTIVE COMBINATIONS

COLLOCATIONS

IDIOMS

Erman, B., & Warren, B. (2000). The idiom principle and the open choice principle. Text & Talk, 20(1), 29–62.
https://doi.org/10.1515/text.1.2000.20.1.29

chase dreams

strong coffee

heavy rain

break promises

swing voters

# kill hope

flood airwaves

read minds

spill secrets

freeze accounts

harbour grudges

fight poverty

throw parties

# read + mind

**читать разум**

V + DObj

**ler mente**

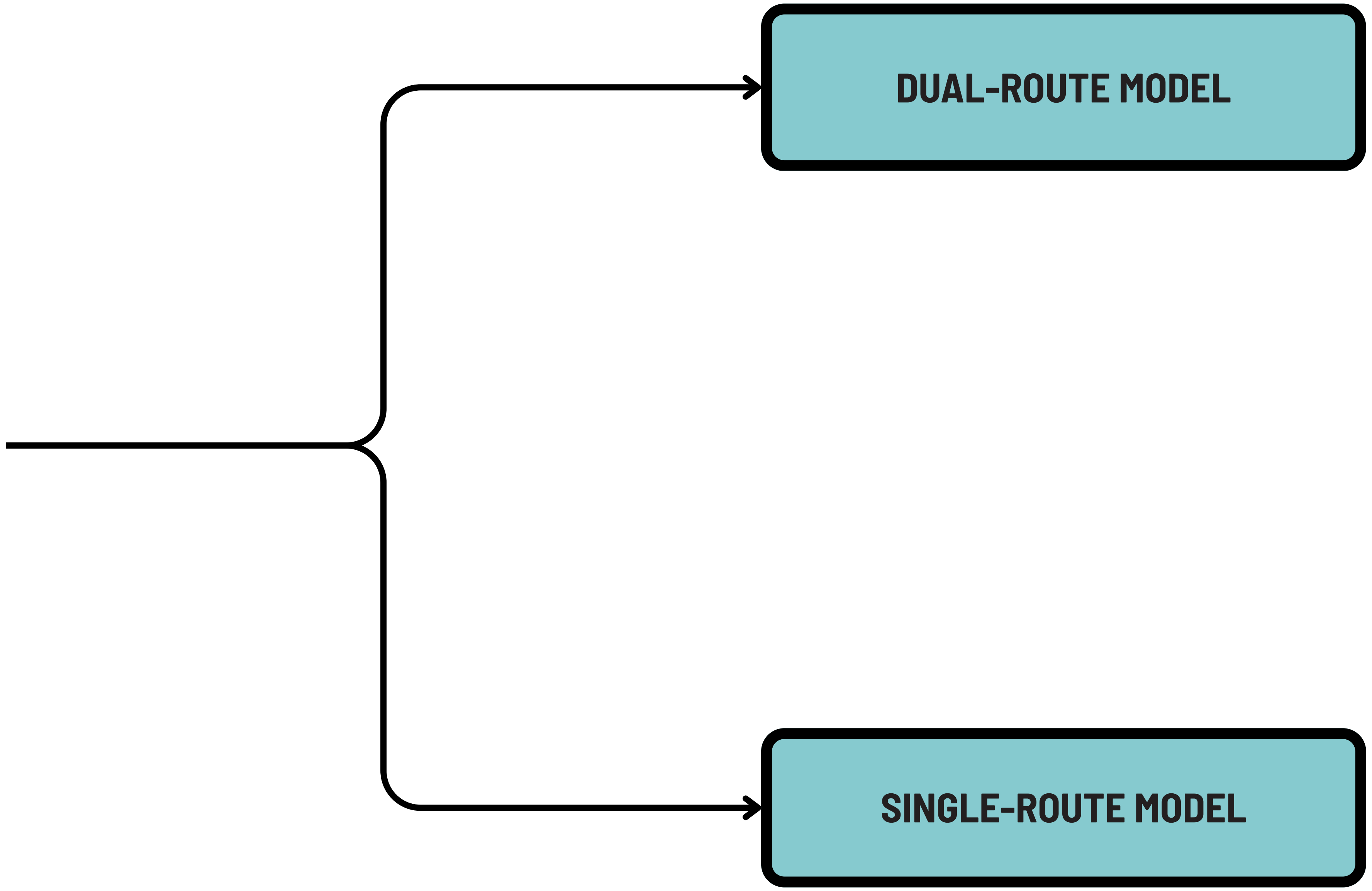
**读心**

**മനസു വായിക്കുക**

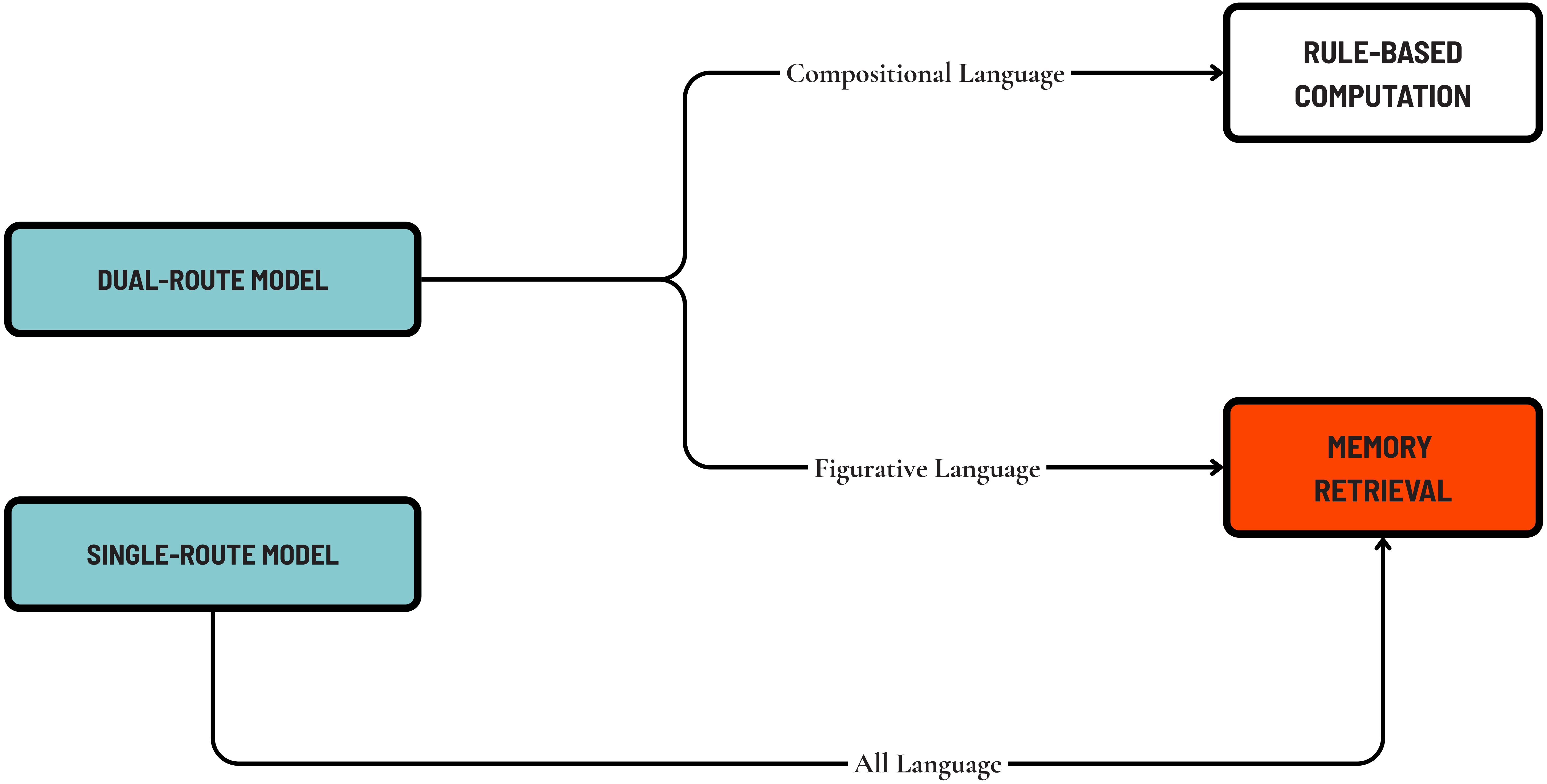
DObj + V

*Why would linguistic structures that are hard to process and hard to learn emerge so ubiquitously in human language?*

# LANGUAGE PROCESSING

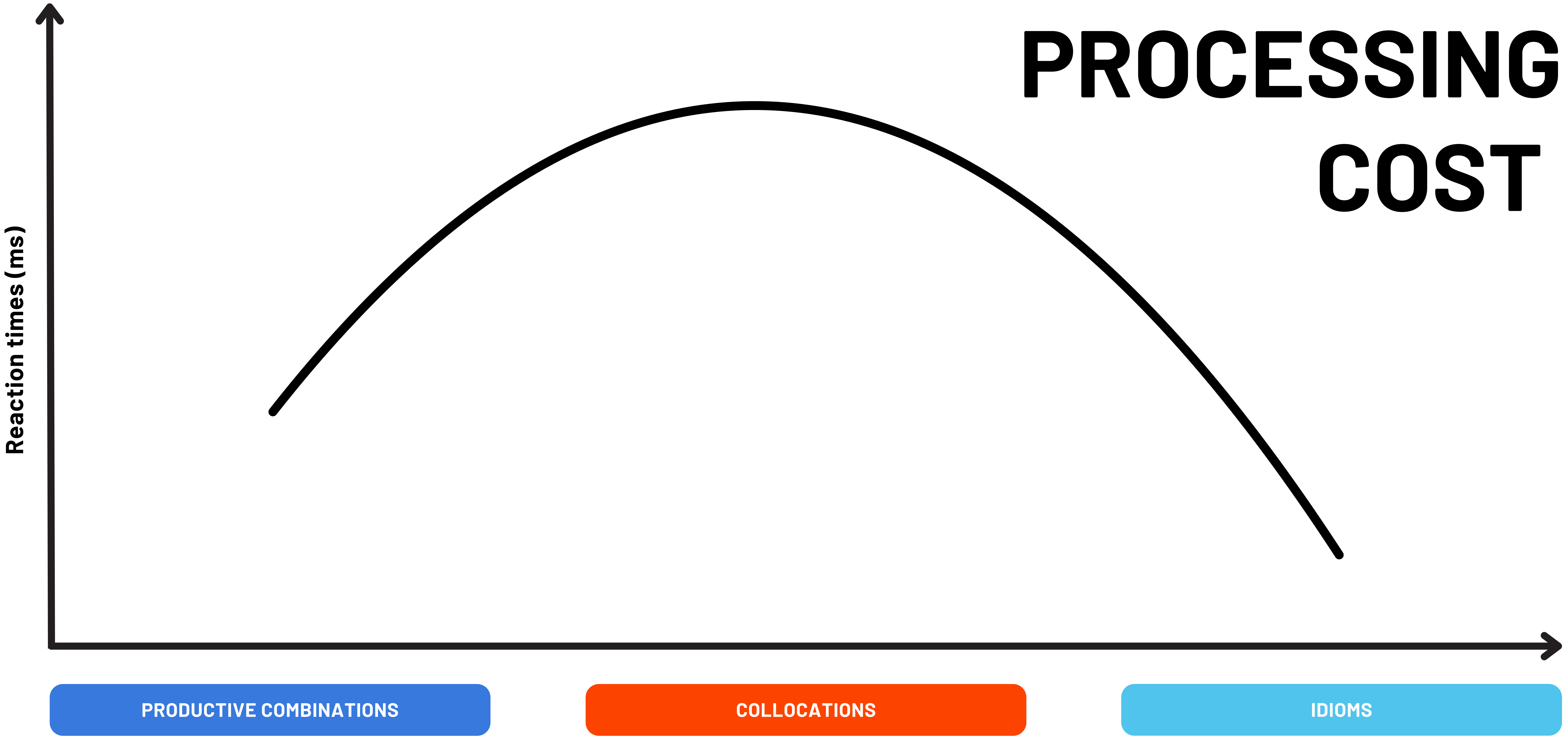



Wray, A. (2002). Formulaic Language and the Lexicon. In Language (Vol. 80). Cambridge University Press.
https://doi.org/10.1353/lan.2004.0209




Wray, A. (2002). Formulaic Language and the Lexicon. In Language (Vol. 80). Cambridge University Press.
https://doi.org/10.1353/lan.2004.0209

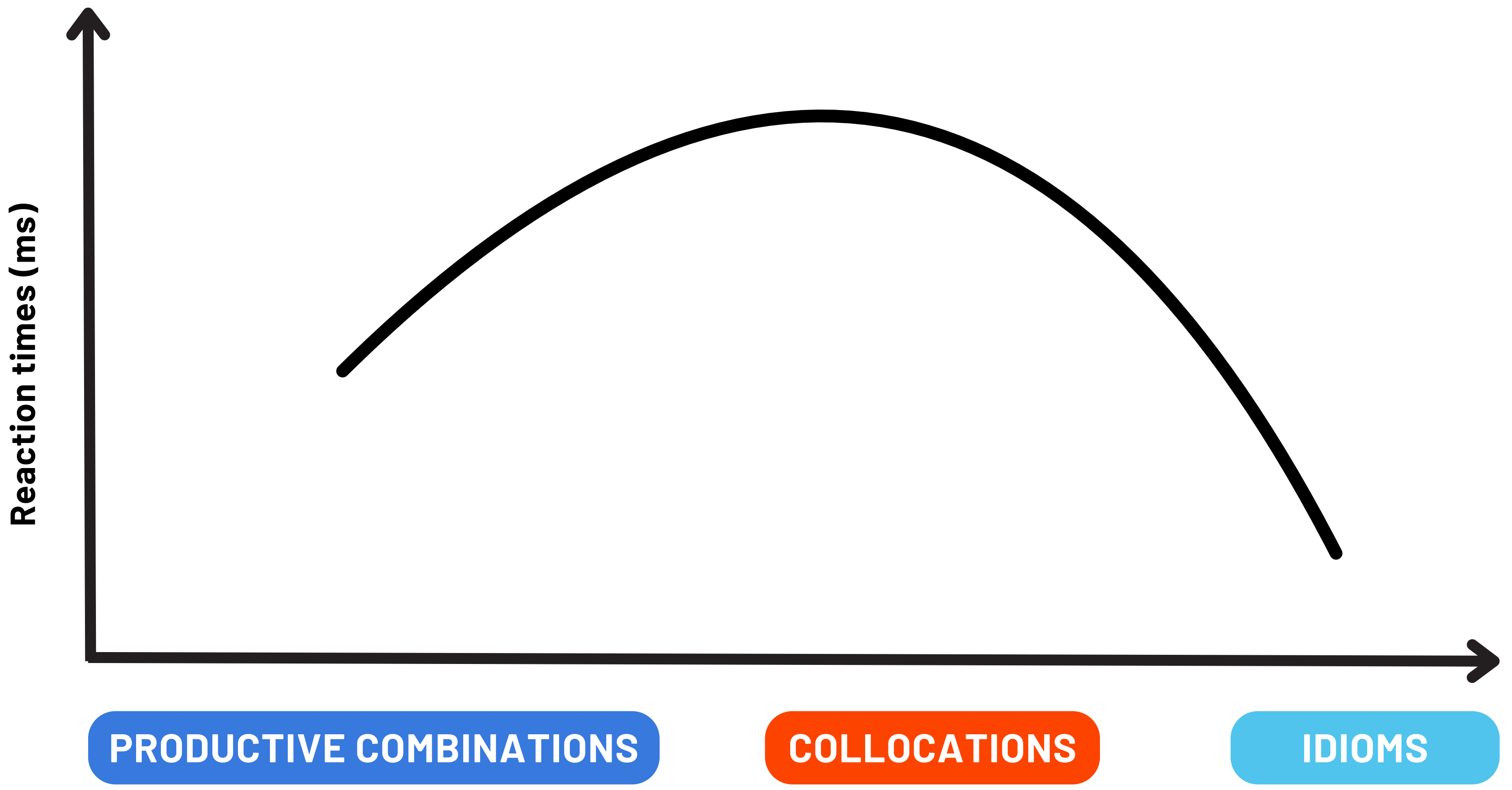
# PROCESSING COST




Imaginary graph. Not to scale.

# AIM

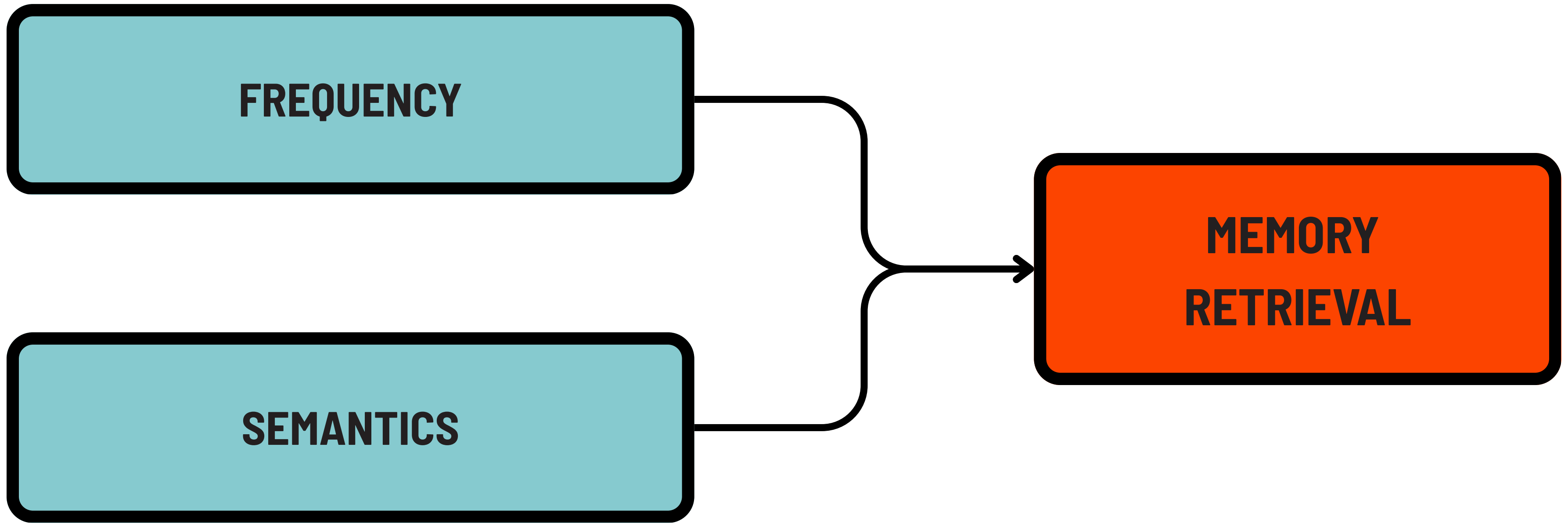
**To test whether pure memory retrieval can account for the processing differences observed in humans**

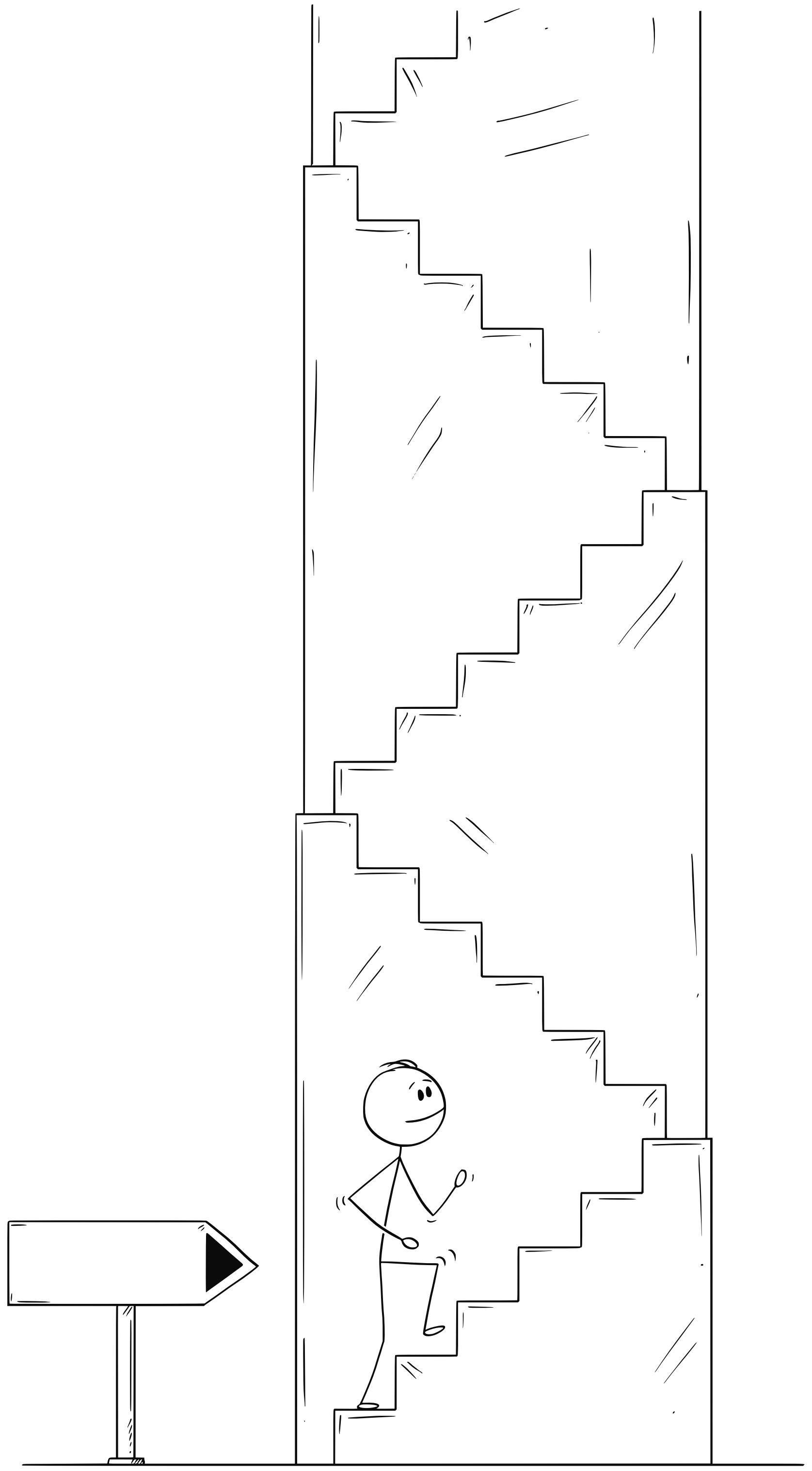


# SPECIFIC RQS

- Do collocations really take longer to process?
- If yes, to what extent can pure memory retrieval account for the processing differences?

# EMERGENT VARIABLES



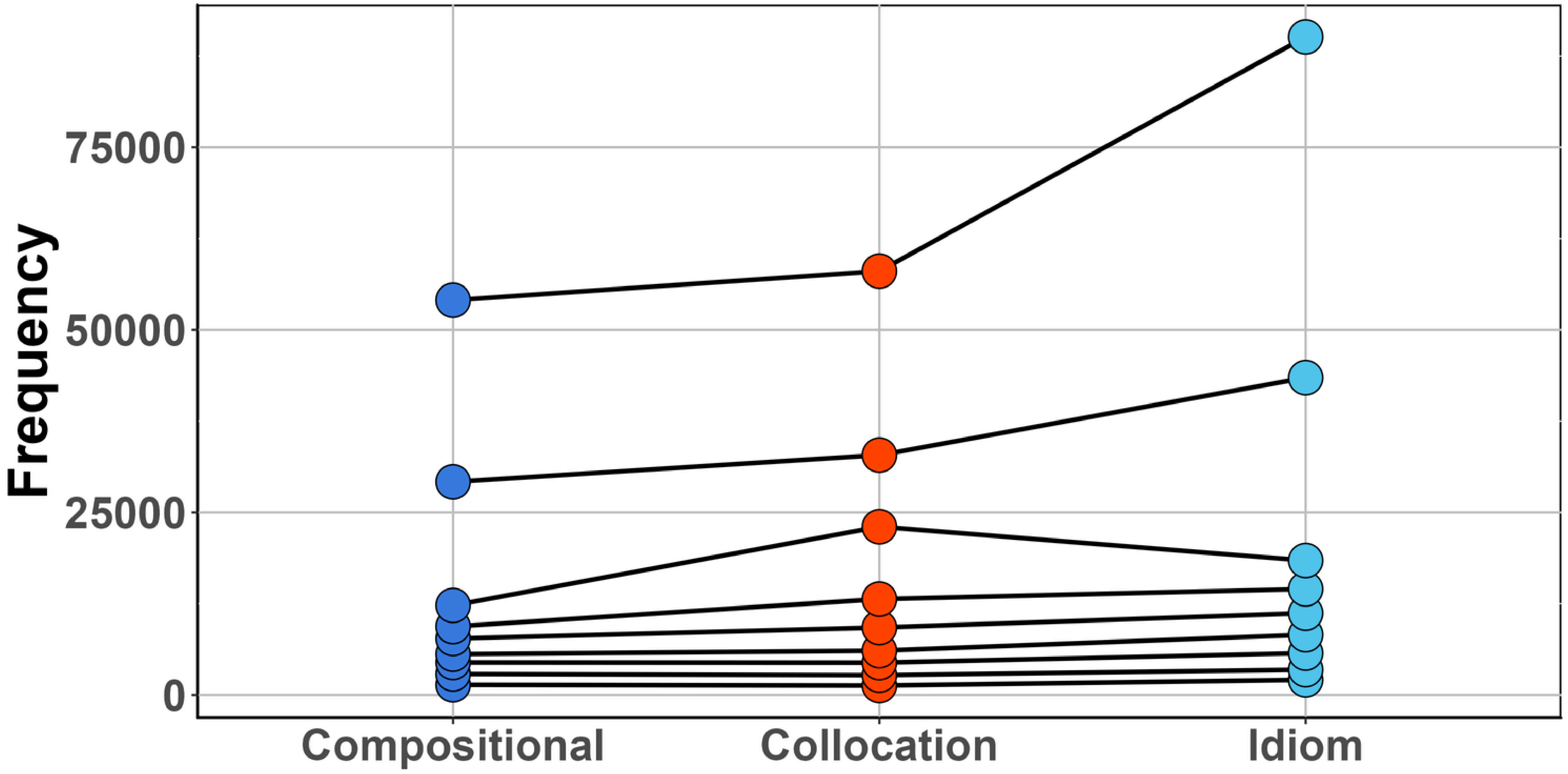
# BEHAVIORAL EXPERIMENT

# #3

A good task elicits behavior that reflects the cognitive process you want to model.

# STIMULI

**240 Verb + Noun Items**
- 80 Productive
- 80 Collocations
- 80 Idioms



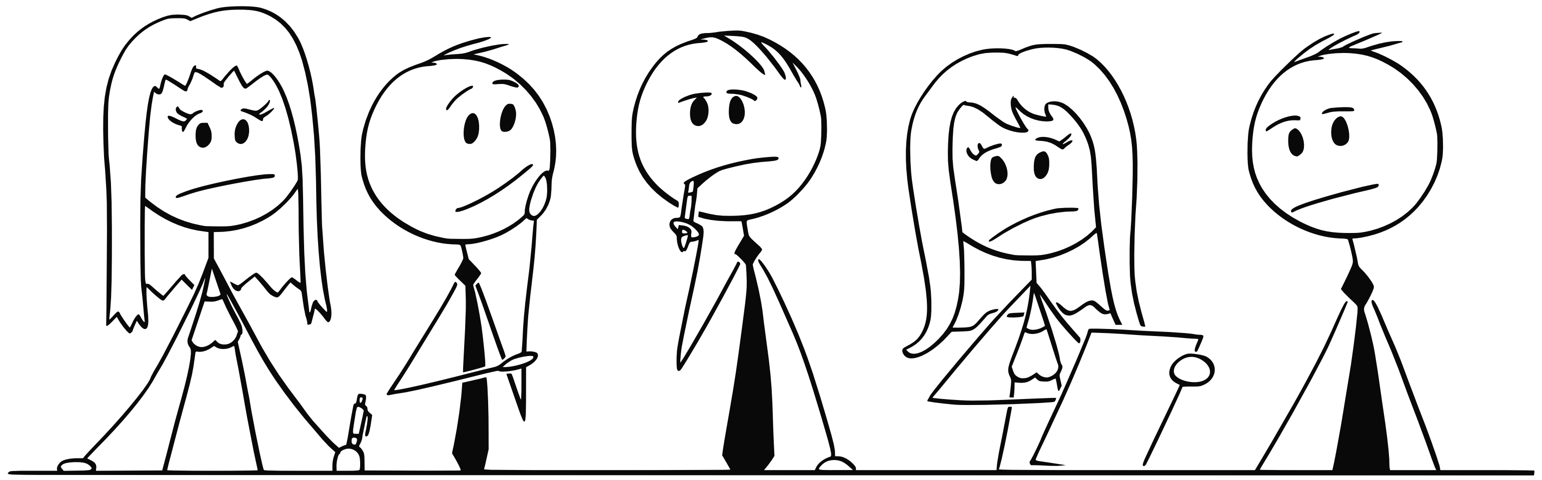

We tried (unsuccessfully) to frequency match items!

# STIMULI

**240 Verb + Noun Items**
- 80 Productive
- 80 Collocations
- 80 Idioms

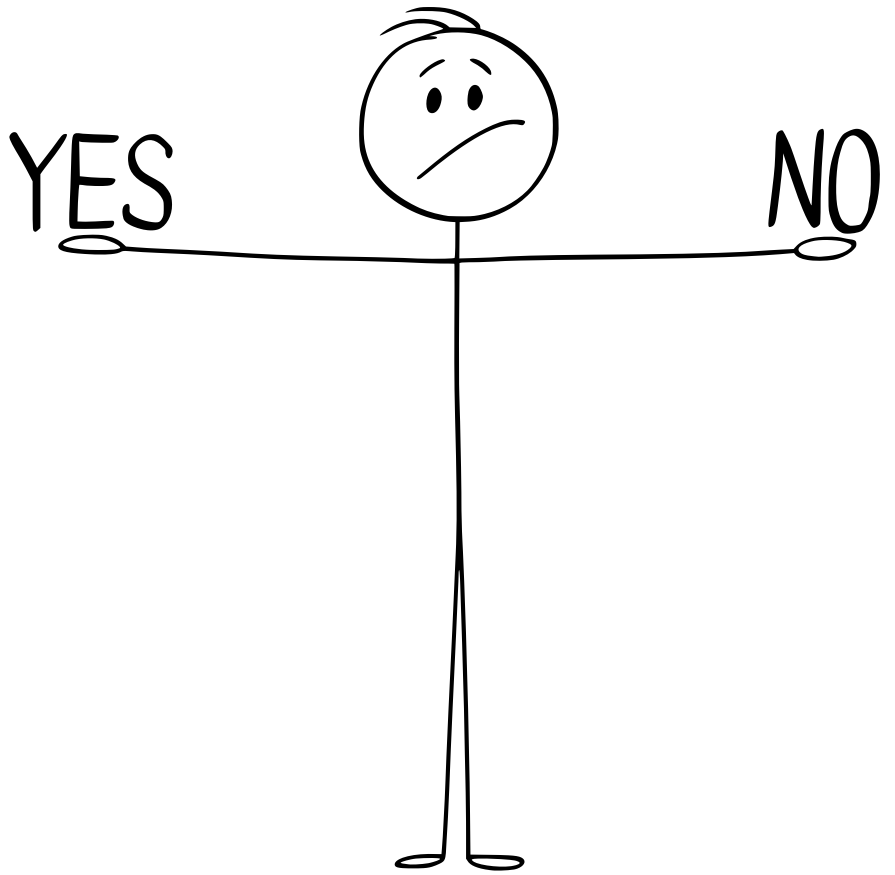
# PARTICIPANTS

**N = 186 humans**
- L1 English speakers
- Mean Age: 38.57 (SD=10.81)
- 117 F, 71 M, 3 NB
- Prolific



# ACCEPTABILITY JUDGEMENT TASK

- **Implicit memory task**
- **Widely used in previous studies**
- Each participant saw 160 items
- Individual randomized order
- 15 second break
- Training set with feedback



# Are the word combinations that appear on the screen acceptable in English?

# eat cakes

ACCEPTABLE

UNACCEPTABLE

## Productive Combination

# eat cakes

✓

ACCEPTABLE

UNACCEPTABLE

# read one’s mind

ACCEPTABLE

UNACCEPTABLE

## Collocation

# read one's mind

ACCEPTABLE

UNACCEPTABLE

# kick the bucket

ACCEPTABLE

UNACCEPTABLE

Idiom

# kick the bucket

ACCEPTABLE

UNACCEPTABLE

# crack cakes

ACCEPTABLE

UNACCEPTABLE

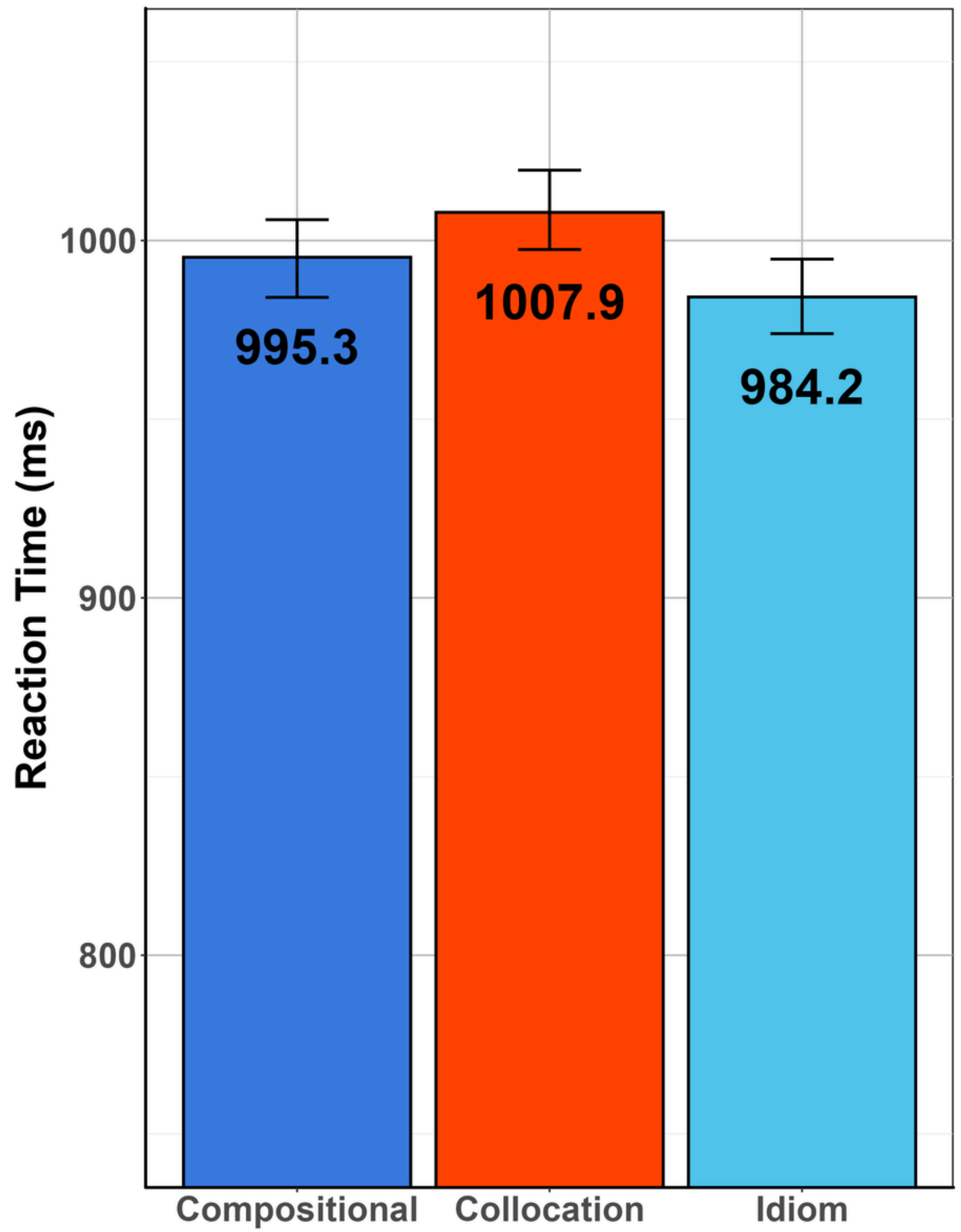
Filler

# crack cakes

ACCEPTABLE

UNACCEPTABLE

# RESULTS




```
RT    Condition + Phrasal Frequency
      + (1 | ID) + (1 | Verb)
```

**Results:**

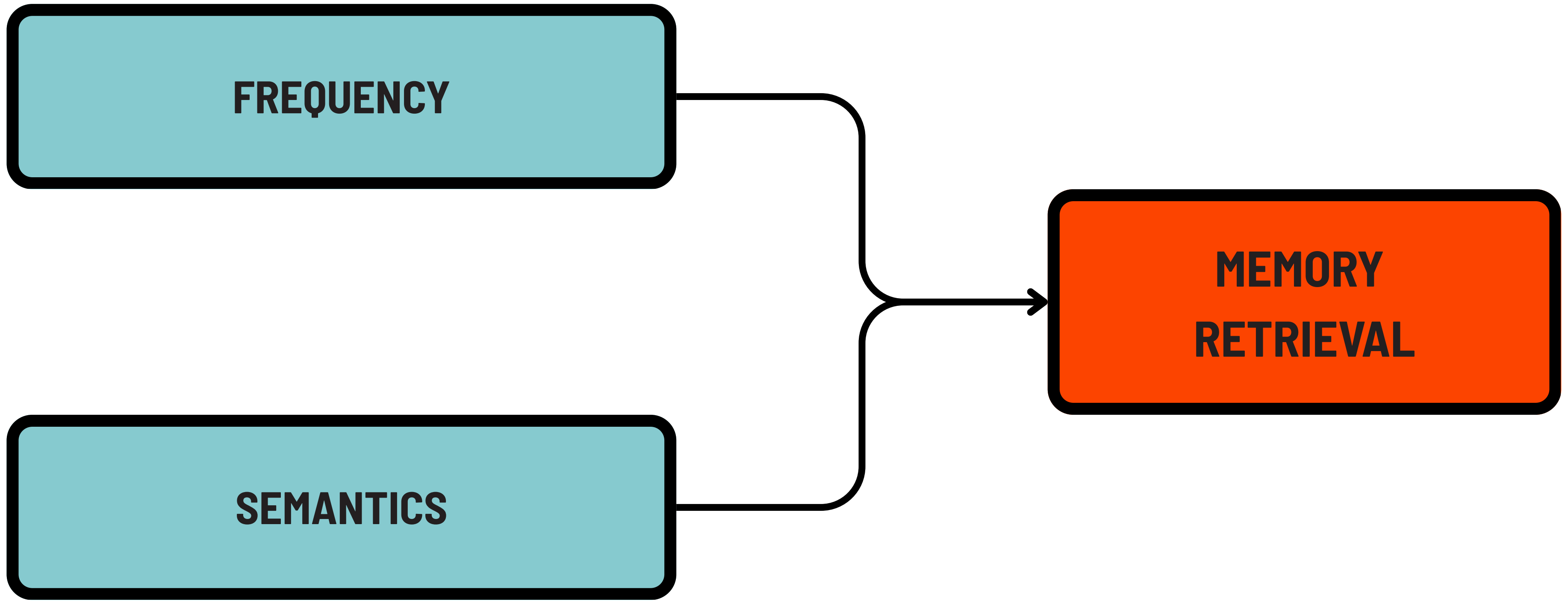
- **Productive-Collocation *****
- **Idiom-Collocation *****
- **Idiom-Productive ***

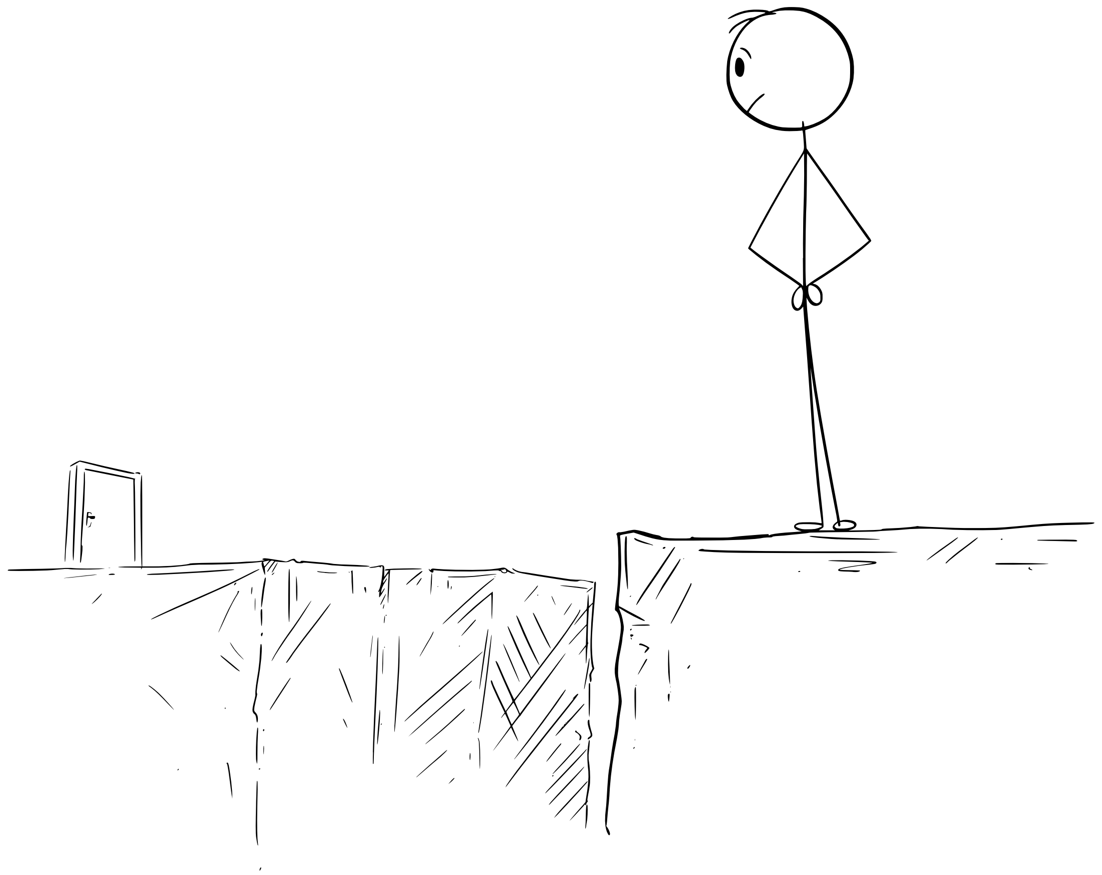
# COMPUTATIONAL MODELLING

# #4

# Modeling means formalizing what participants might be doing cognitively.

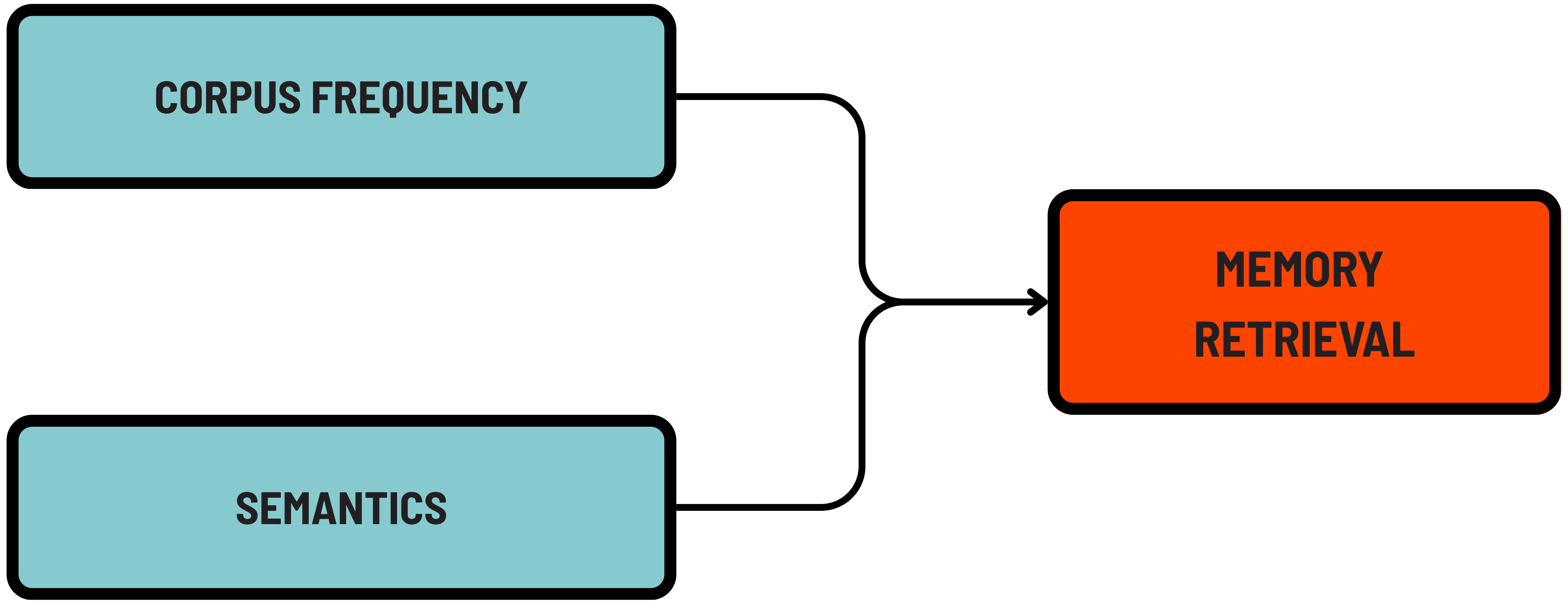
# DESIDERATA



```
RT    Condition + Phrasal Frequency
      + (1 | ID) + (1 | Verb)
```

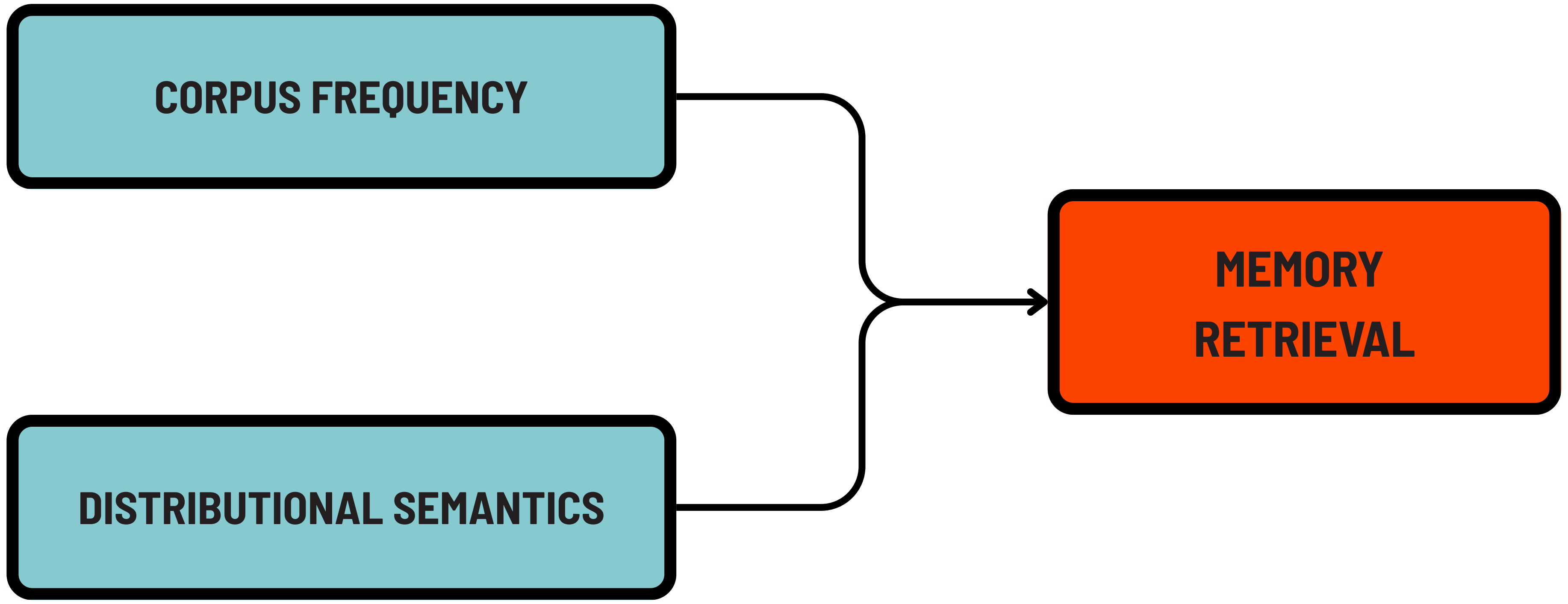
# MODELLING FREQUENCY



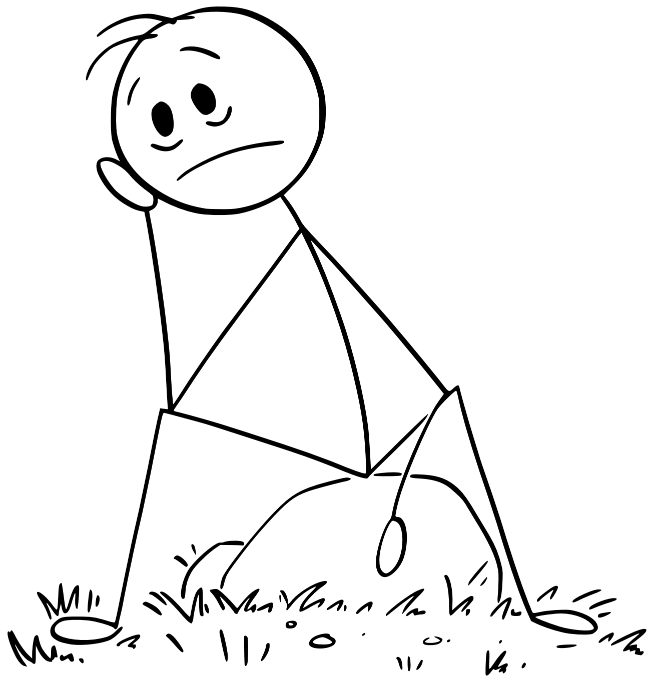

RT Condition + Phrasal Frequency

+ (1 | ID) + (1 | Verb)

# MODELLING SEMANTICS

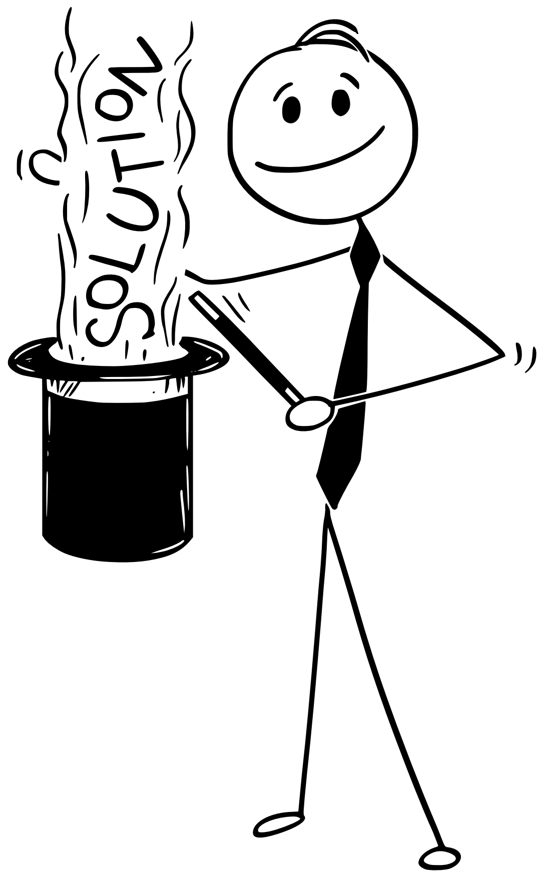



```
RT    Condition + Phrasal Frequency
      + (1 | ID) + (1 | Verb)
```



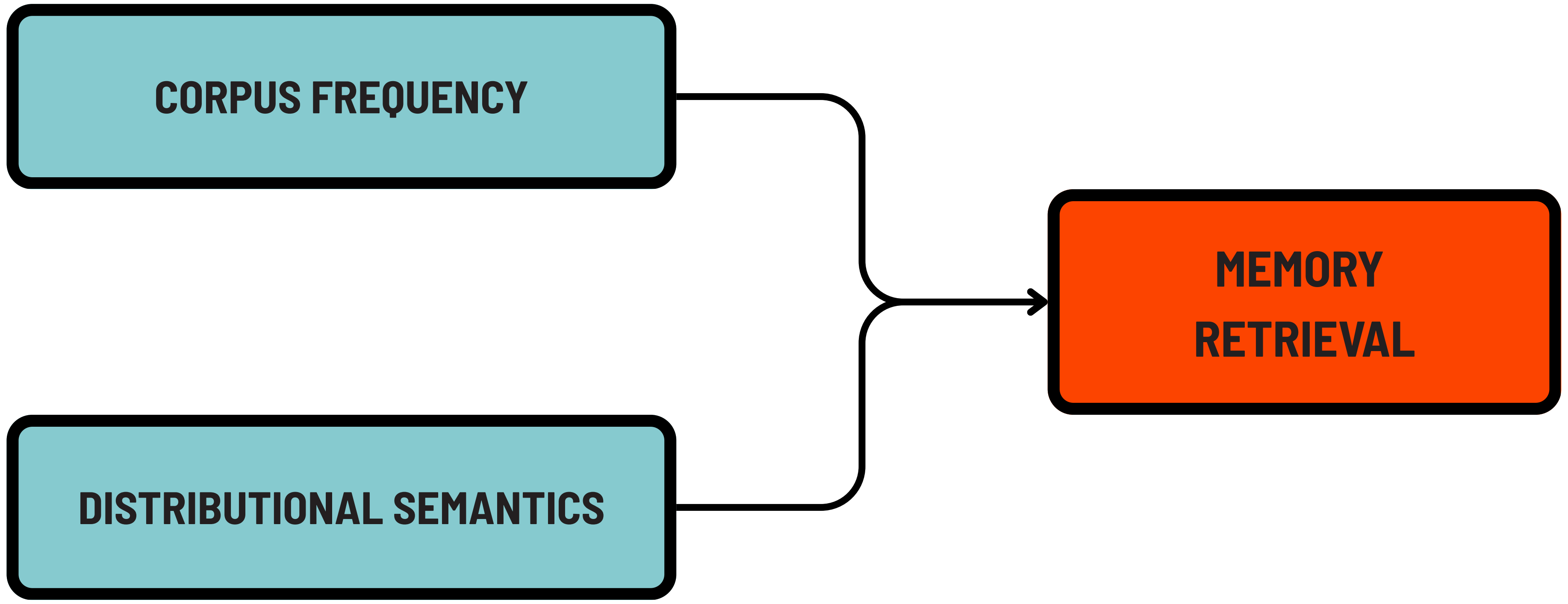
# EMBEDDINGS

- **Contextual embeddings**
  - Collect 100 context sentences containing word combination from the corpus
  - Embed sentences using SentenceBERT
  - Pick out embeddings for verb and noun from each sentence
  - Average verb and noun embeddings across sentences, separately
- **Item embedding = concatenated embeddings of verb and noun**

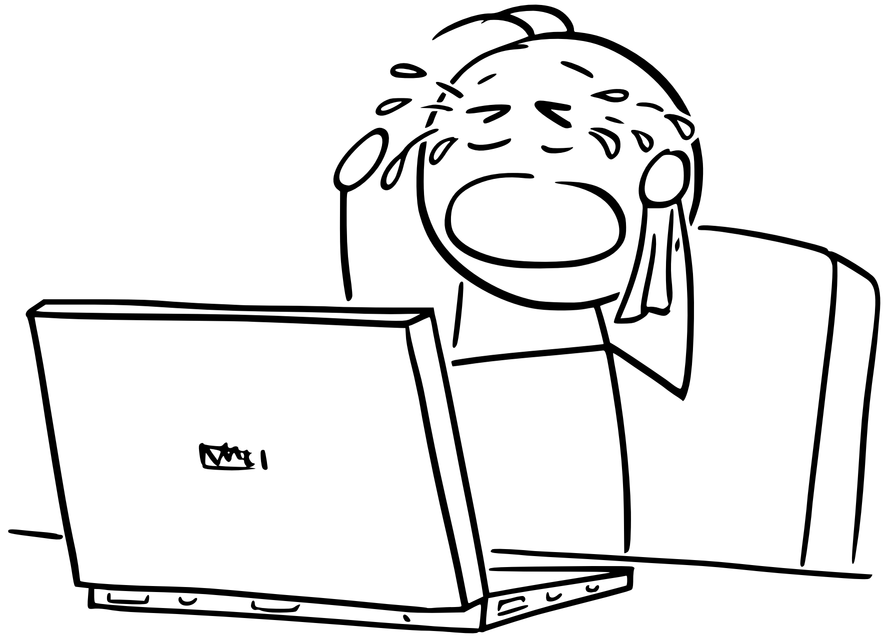



Vulić, I., Ponti, E. M., Litschko, R., Glavaš, G., & Korhonen, A. (2020). Probing Pretrained Language Models for Lexical Semantics. In B. Webber, T. Cohn, Y. He, & Y. Liu (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 7222–7240). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.586

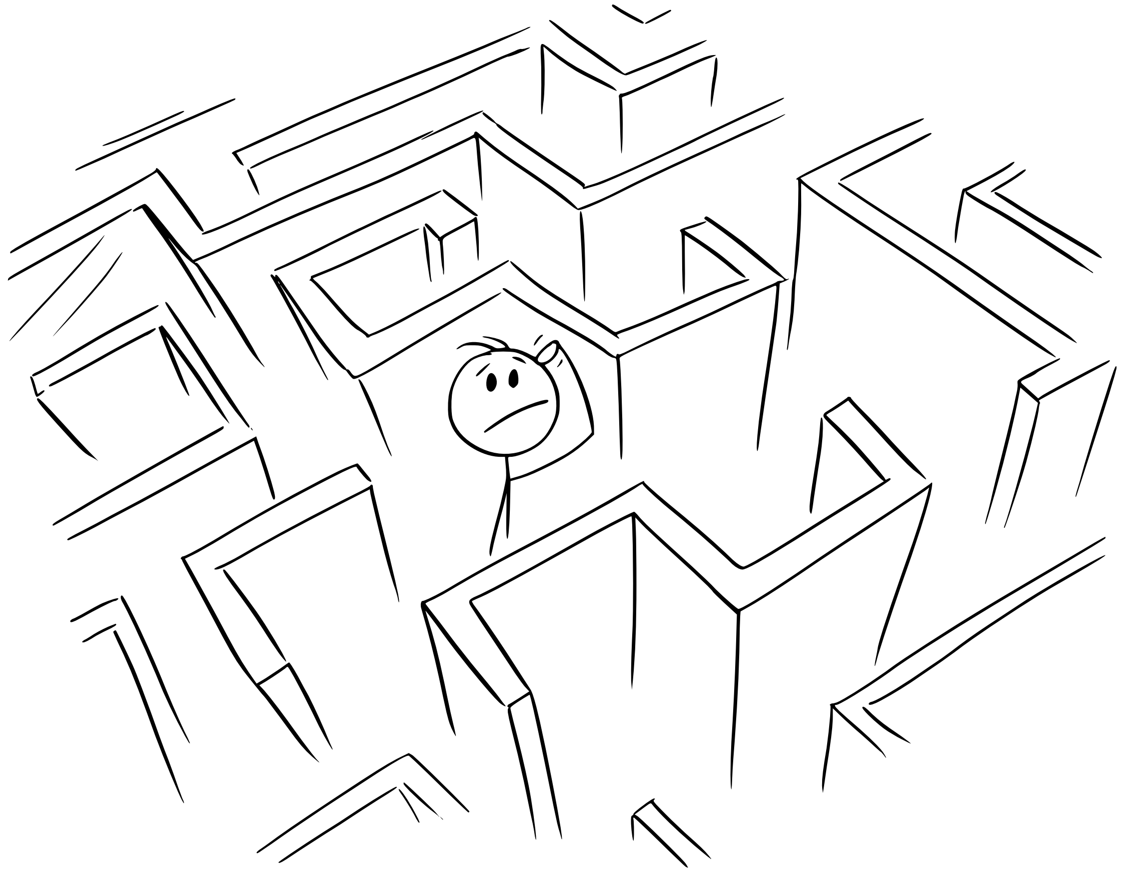
# MODELLING MEMORY



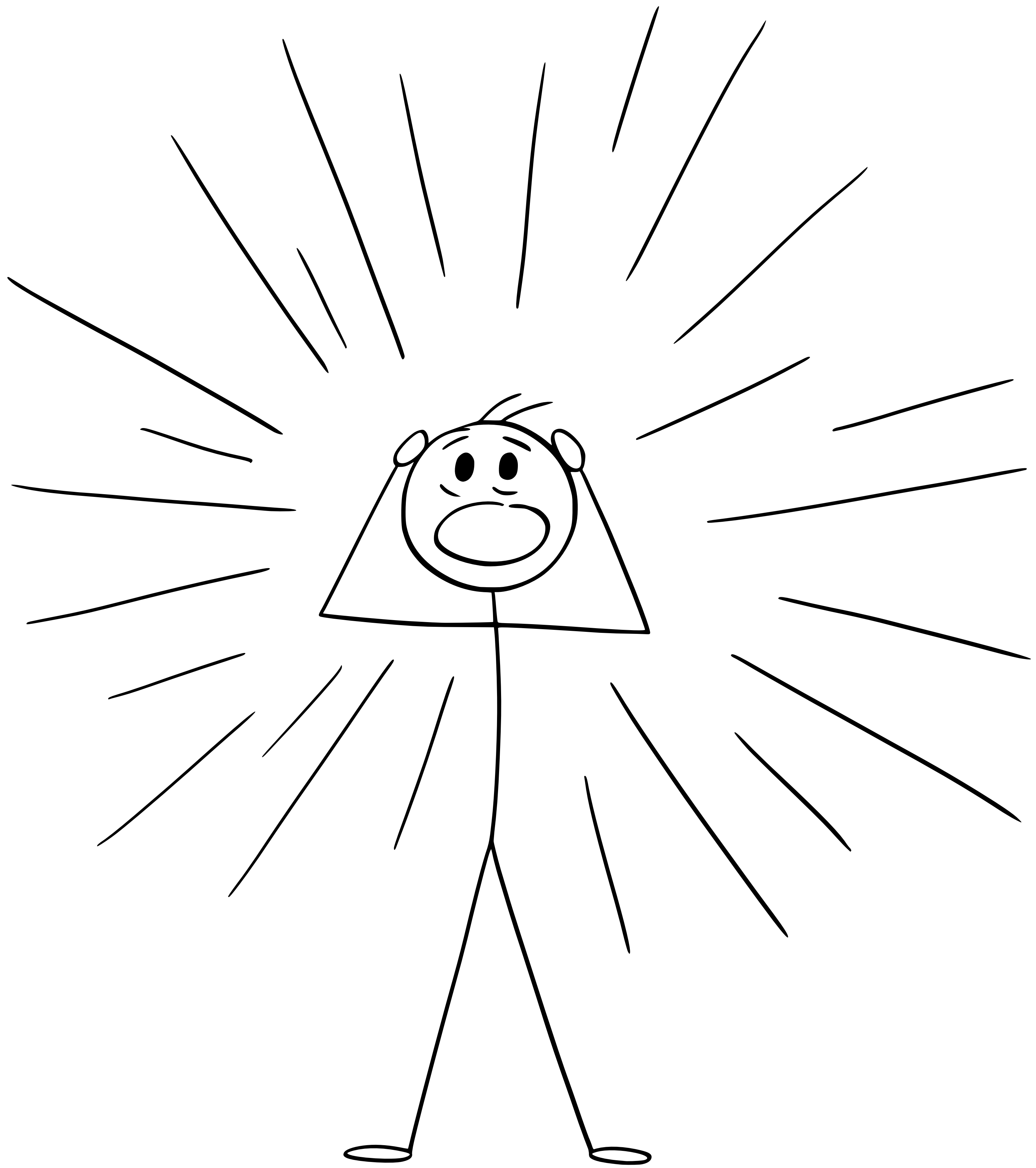

```
RT    Condition + Phrasal Frequency
      + (1 | ID) + (1 | Verb)
```



# SEEMS LIKE EVERYONE AND THEIR MOM HAS BUILT A MEMORY MODEL!



| Category | Models | Retrieval Mechanism | Major Limitations |
|---|---|---|---|
| **Trace-Based / Global Matching** | *MINERVA 2, REM, SAM* | Global similarity; sampling across stored traces | Ad hoc similarity metrics and scaling parameters |
| **Context-Based / Temporal** | *TCM, CMR, eCMR* | Reinstating prior context or temporal cues | Too many free parameters! |
| **Connectionist** | *Hopfield, HRR, CLS* | Pattern completion in distributed representations | Hard to interpret opaque retrieval dynamics |
| **Probabilistic / Bayesian** | *BART, Bayesian Memory* | Probabilistic inference under uncertainty | Only explains “optimal” behavior<br>Computationally expensive; analytically intractable |
| **Neural / AI-Inspired** | *Modern Hopfield, MANNs, RAG, NTM* | Attention or key–value lookup in differentiable memory | Not cognitively plausible! |
| **Hybrid / Cognitive Architectures** | *ACT-R, Soar, LEABRA, CHREST* | Combine symbolic control with subsymbolic retrieval mechanisms | Complex and parameter-heavy<br>Integrate multiple mechanisms without clear theoretical boundaries |

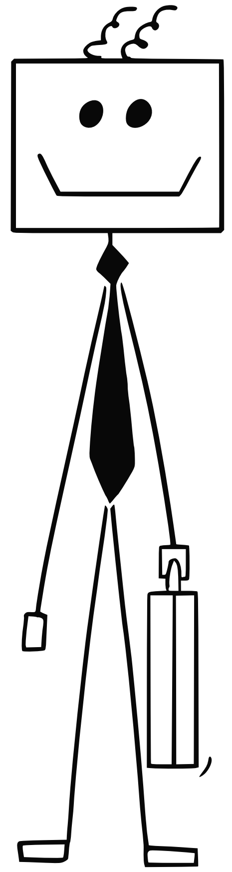


# #5

The goal isn't a perfect fit, but rather a transparent test of your theoretical assumptions.

# MINERVA2

- A **parsimonious**, **tractable** model of episodic memory
- **3 core assumptions**:
  - each experience leaves a trace of features
  - similar inputs → similar traces
  - retrieval driven by global similarity matching
- **Integrates episodic** and **semantic memory** through parallel, item-specific retrieval



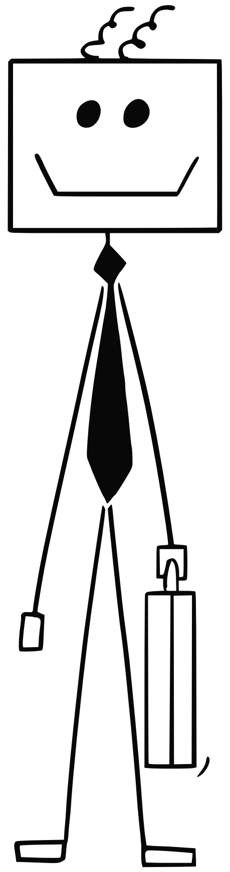
**TLDR: offers predictive precision while remaining computationally efficient**

Hintzman, D. L. (1984). MINERVA 2: A simulation model of human memory. Behavior Research Methods, Instruments, & Computers, 16(2), 96–101.
https://doi.org/10.3758/BF03202365

# MINERVA2

- **Psychologically grounded**
- **Makes testable predictions**
- Successfully applied to diverse cognitive domains, including language:
  - Frequency judgments (Hintzman, 1988)
  - False memory (Arndt & Hirshman, 1998)
  - **Artificial grammar learning** (Jamieson & Mewhort, 2009)
  - **Metaphor recognition** (Reid & Jamieson, 2023)

**TLDR: is empirically consistent and falsifiable**



# MINERVA2

$$s = sim(p, \boldsymbol{M})$$

$$a_\tau = s^\tau \, sign(s)$$

$$e_\tau = a_\tau \boldsymbol{M}$$

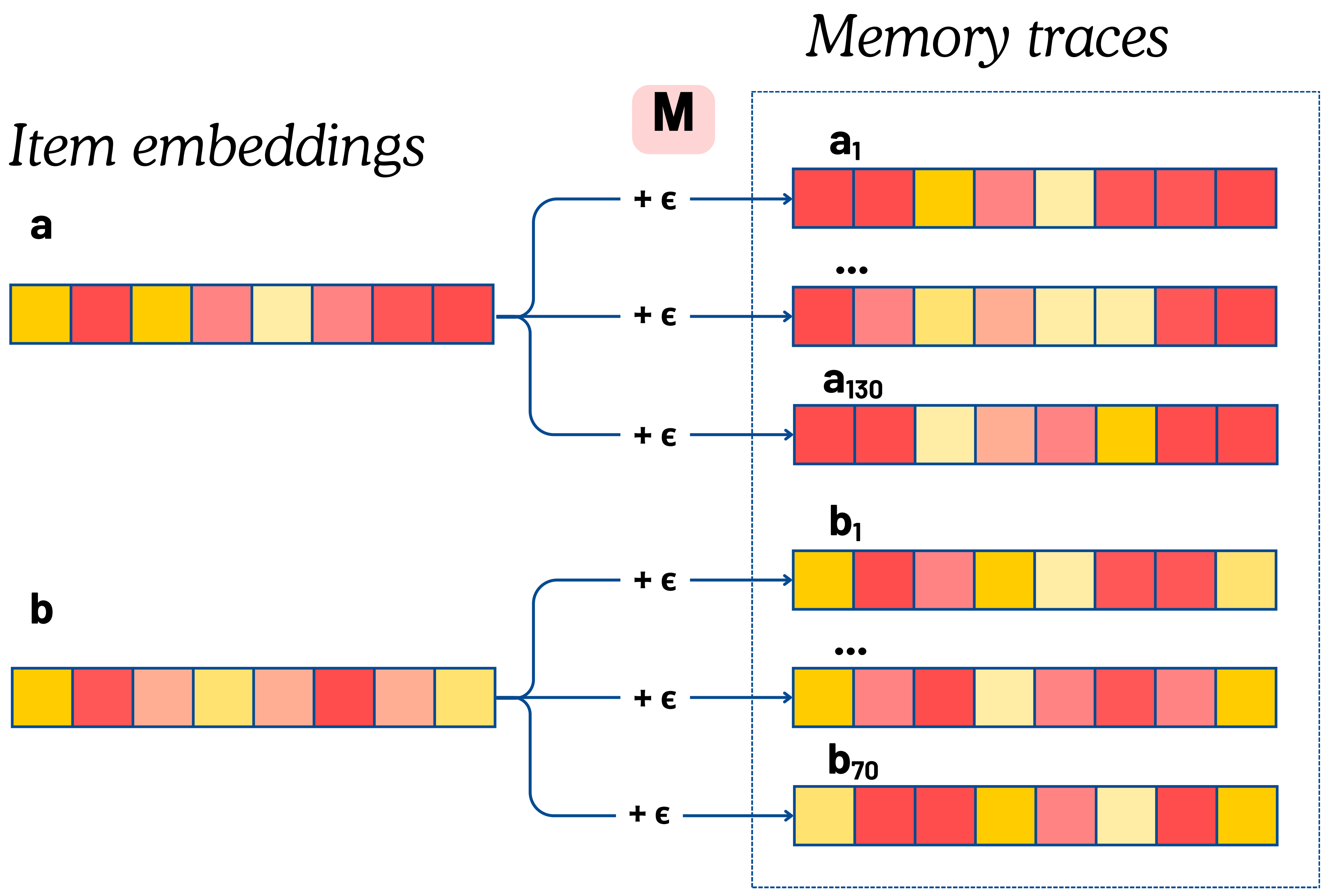
$$f_\tau = sim(e_\tau , p)$$

Nick Reid, J., & Jamieson, R. K. (2023). True and false recognition in MINERVA 2: Extension to sentences and metaphors. Journal of Memory and Language, 129, 104397. https://doi.org/10.1016/j.jml.2022.104397

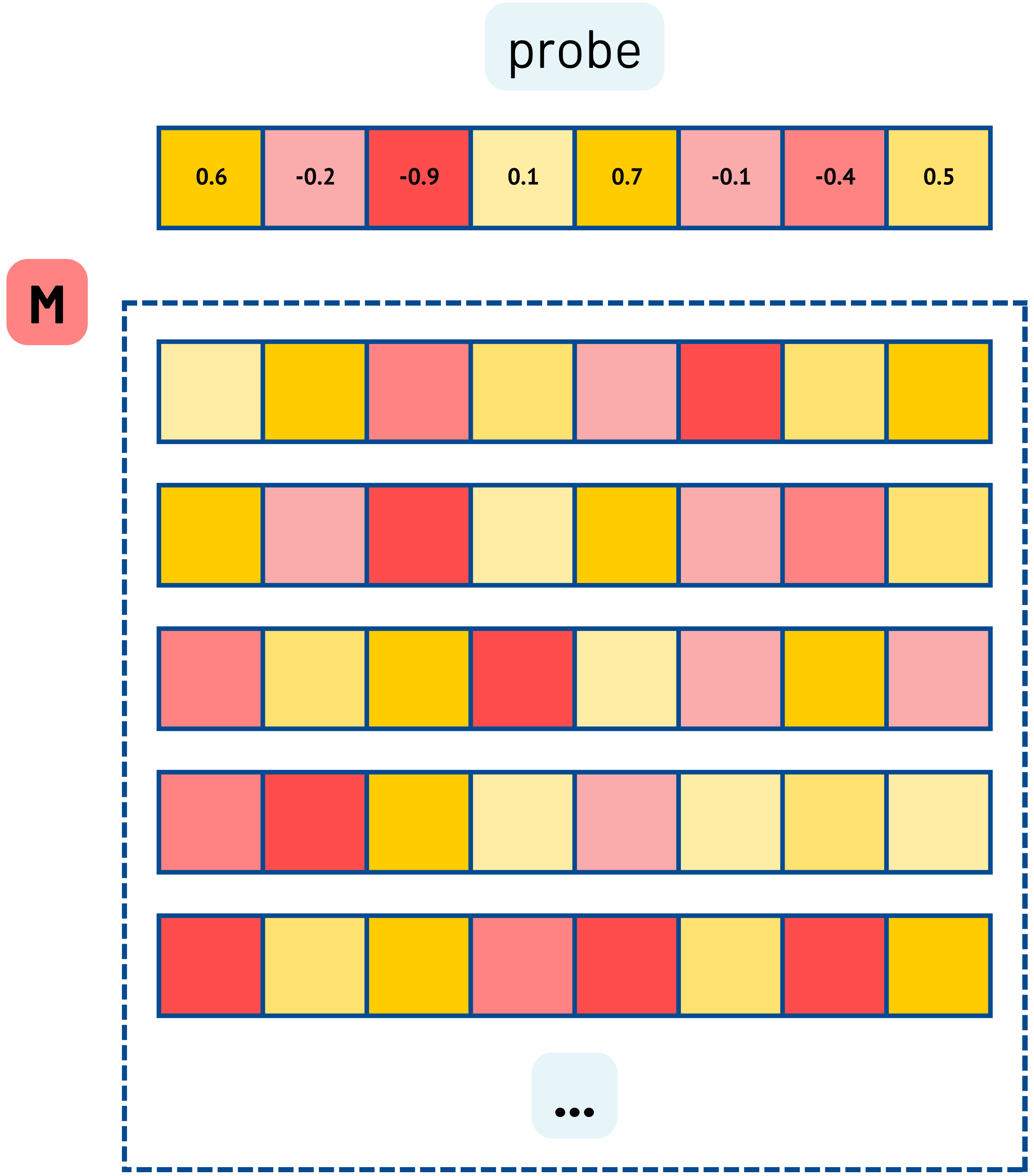
# EXPOSURE + STORAGE

Sample item embeddings into MINERVA memory

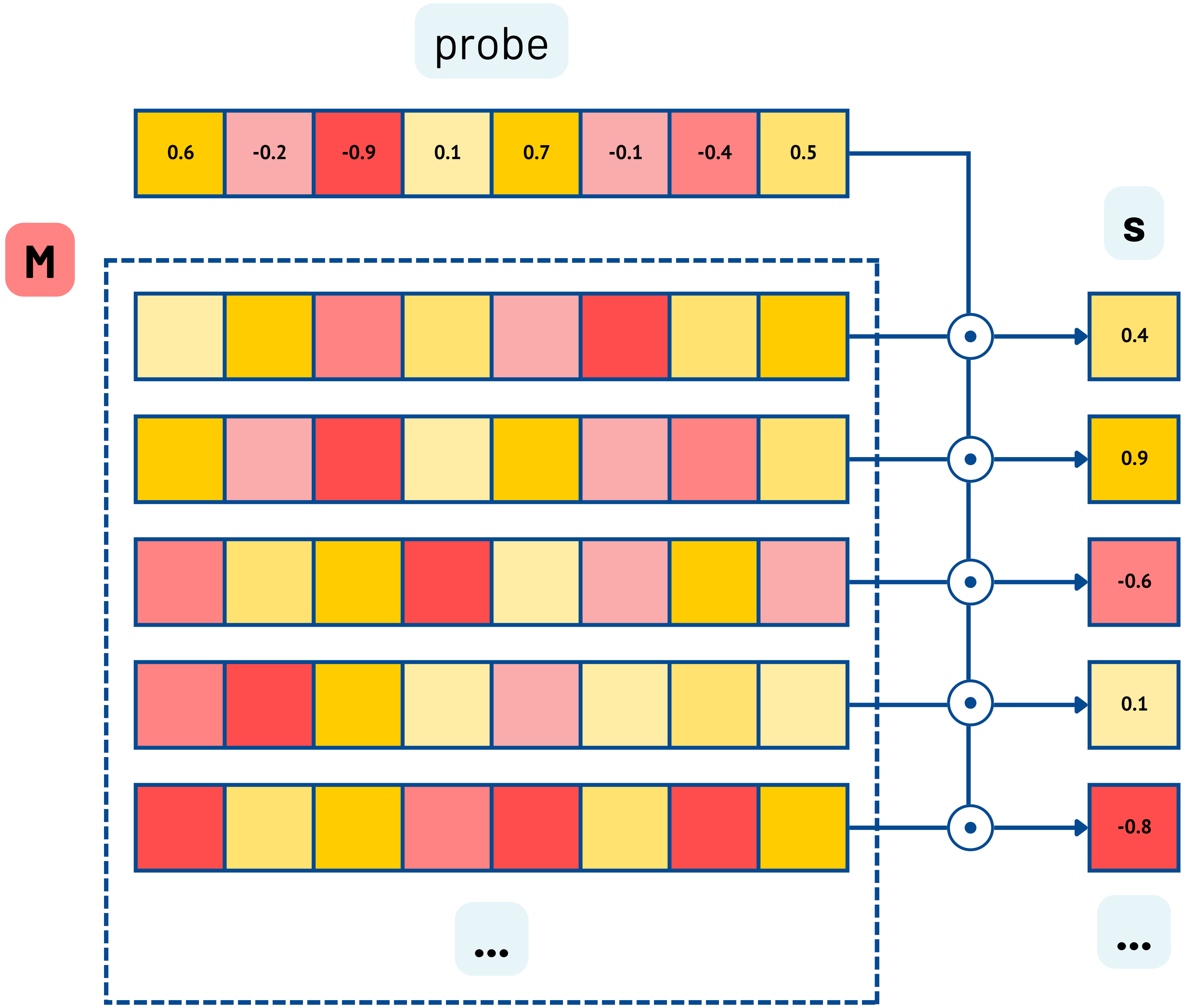
- according to corpus frequency (n)
- Each row is a d-dimensional memory trace.
- $M \in \mathbb{R}^{n \times d}$
- Randomly noise each dimension of each embedding
    - Simulates forgetting



$$s = sim(p, \boldsymbol{M})$$



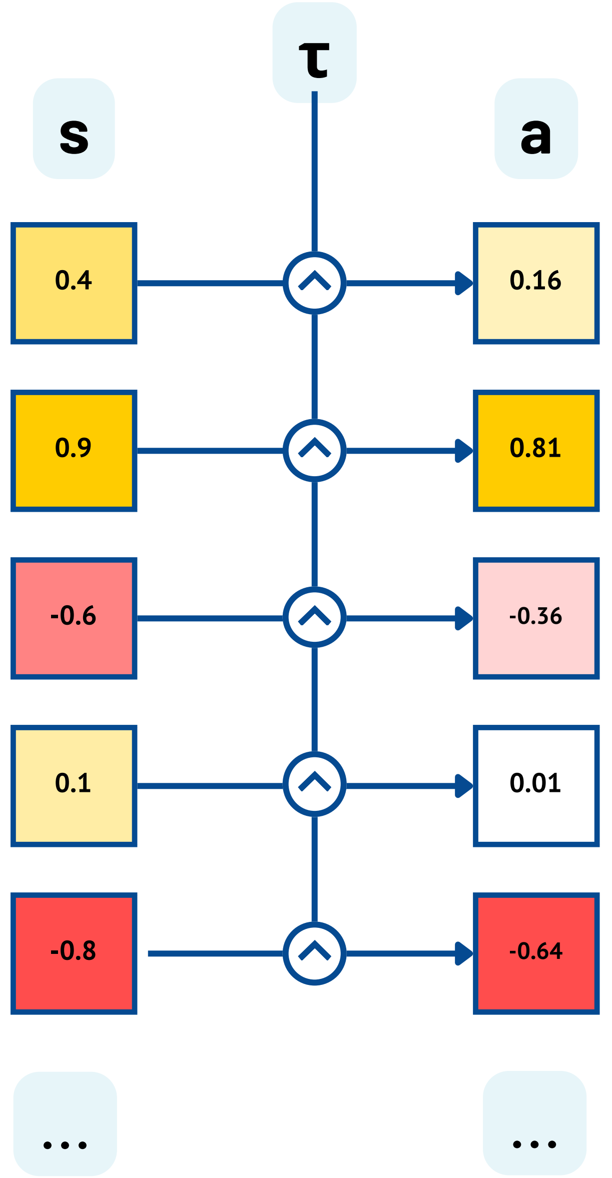
$$s = sim(p, \boldsymbol{M})$$




Compute similarity between the probe and each item in memory → activation

- Probe: $\boldsymbol{p} \in \mathbb{R}^d$
- represents current input

$$a_\tau = s^\tau \, sign(s)$$

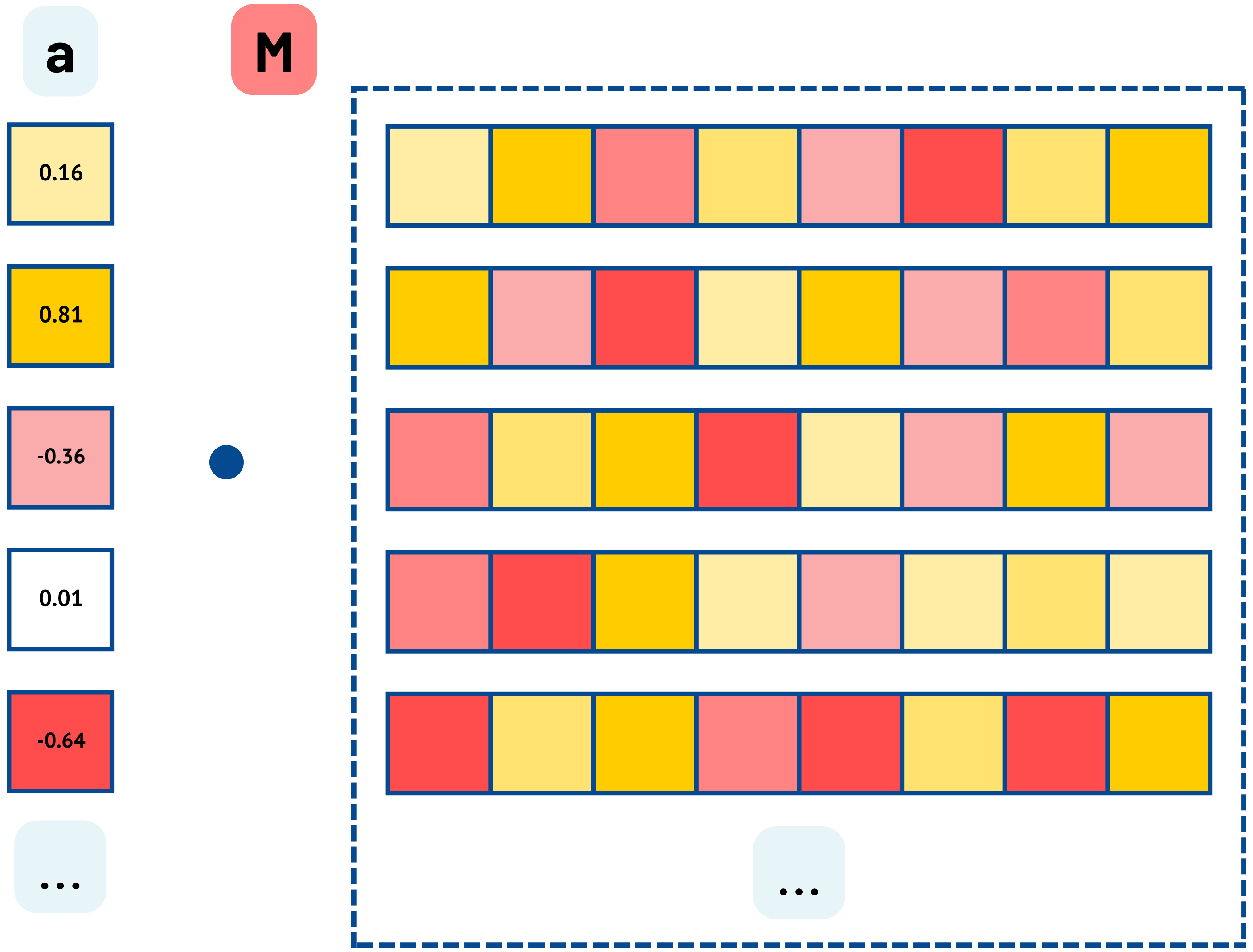


τ-accentuated similarities

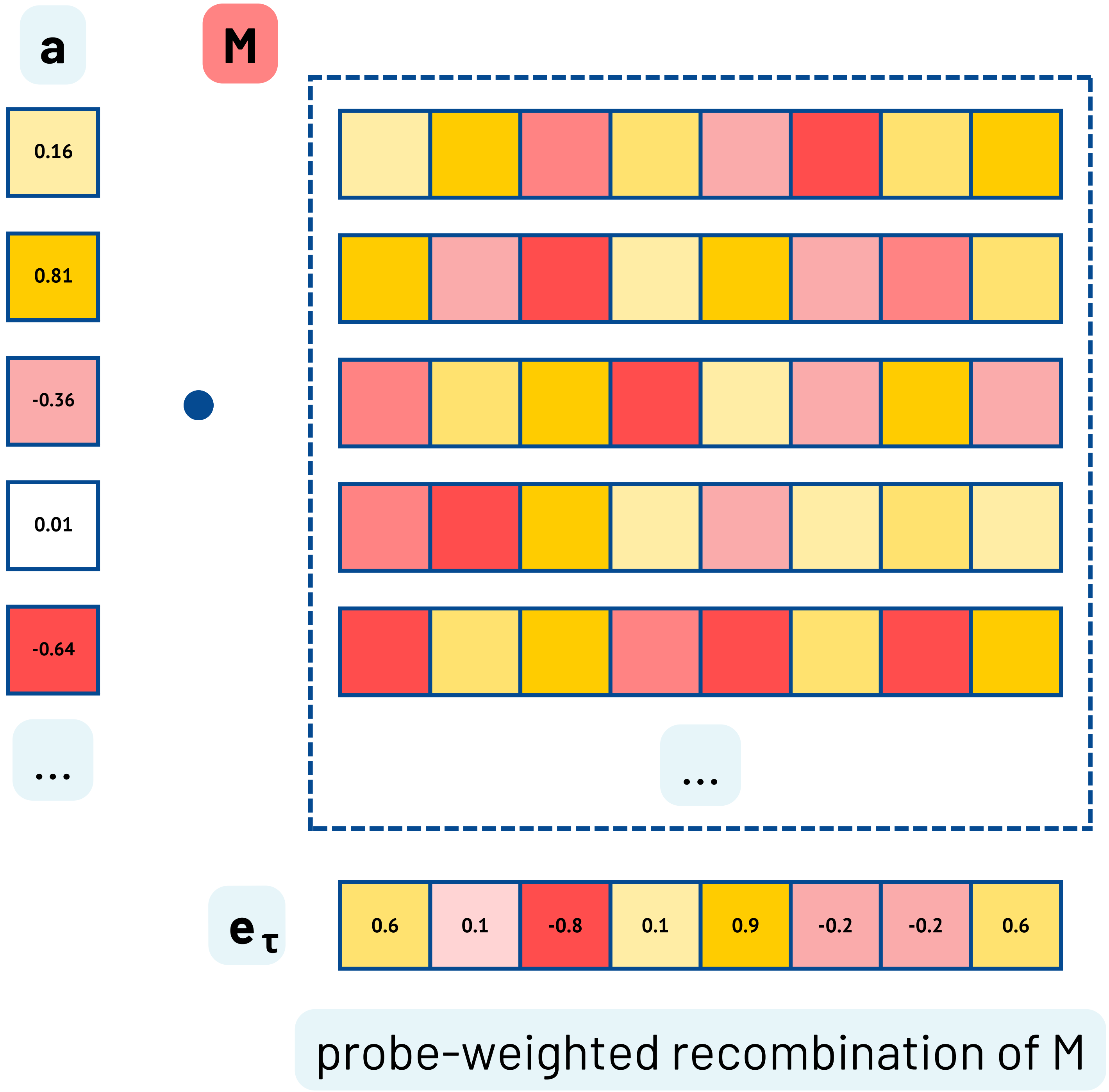
- Sharpen activations by exponentiating them by Tau
- High magnitudes stay high, low magnitudes get decayed
- Temporal index, i.e., no. of iterations required to reach a threshold

$$e_\tau = a_\tau \mathbf{M}$$

Average all items in memory according to the sharpened activation of each



$$e_\tau = a_\tau \mathbf{M}$$



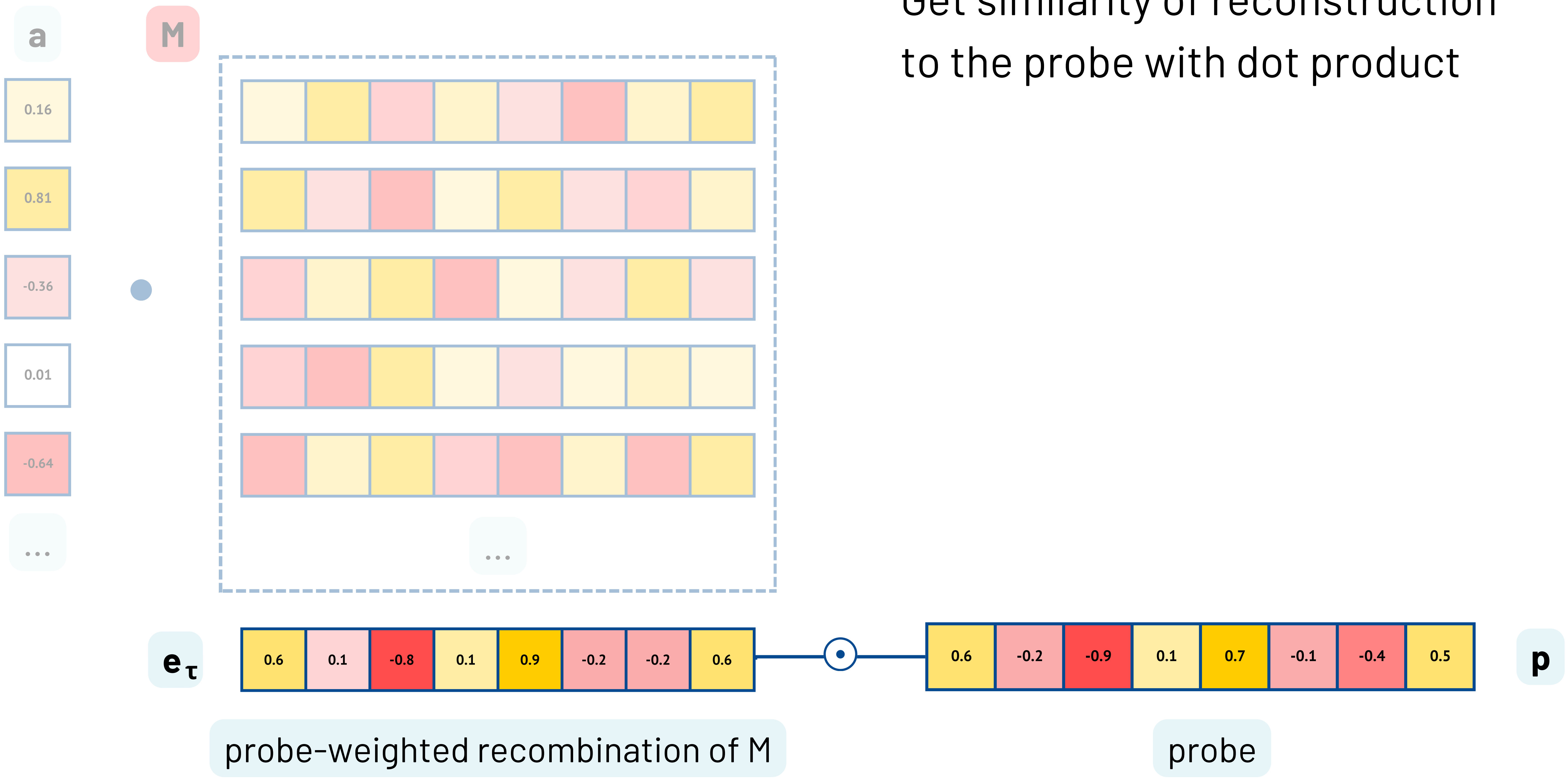

Average all items in memory according to the sharpened activation of each

- Uses the memory to reconstruct the probe
- i.e., reconstruct probe from past traces

$$e_\tau = a_\tau \mathbf{M}$$

$$f_\tau = sim(e_\tau , p))$$

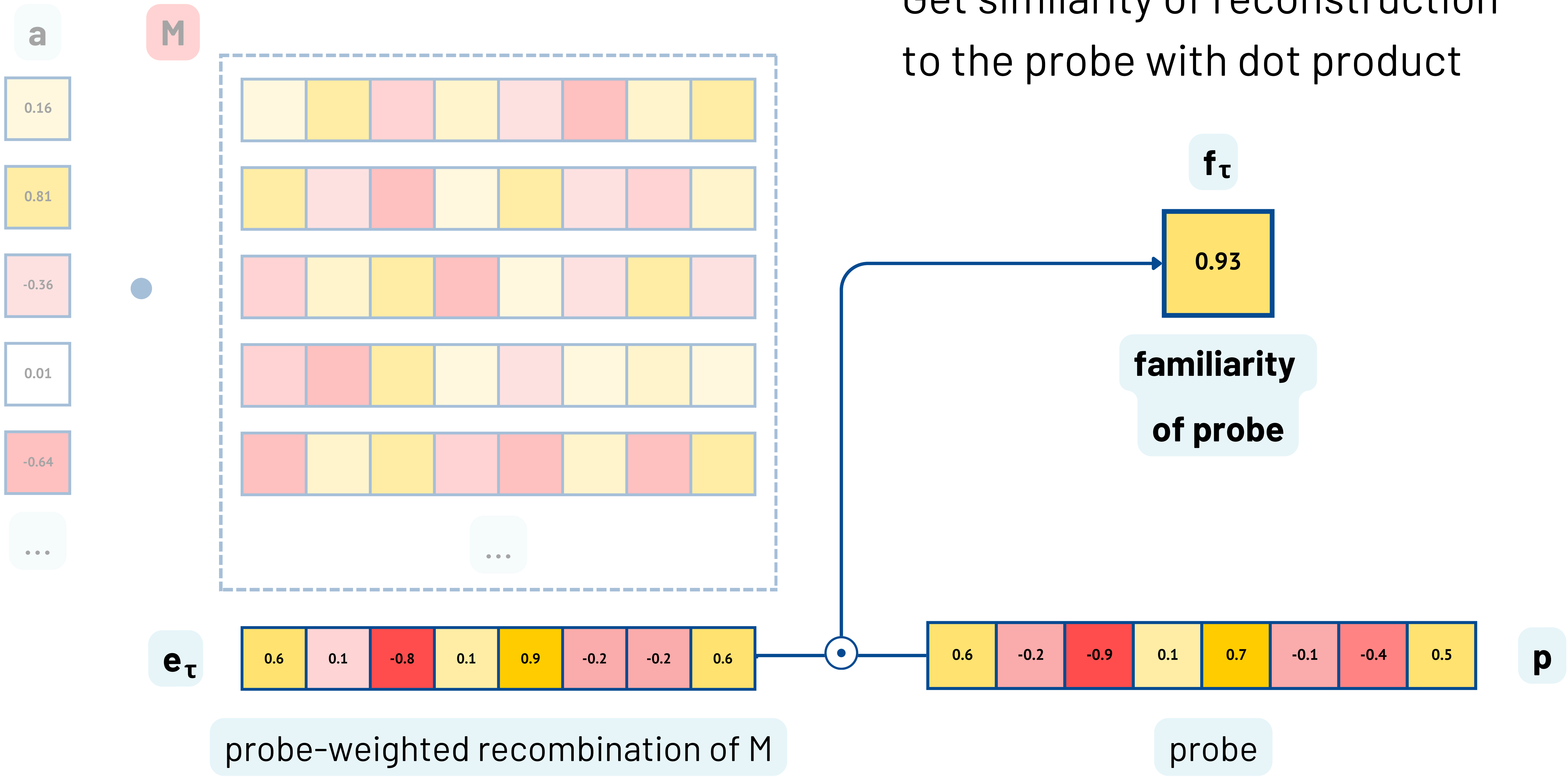
Get similarity of reconstruction to the probe with dot product

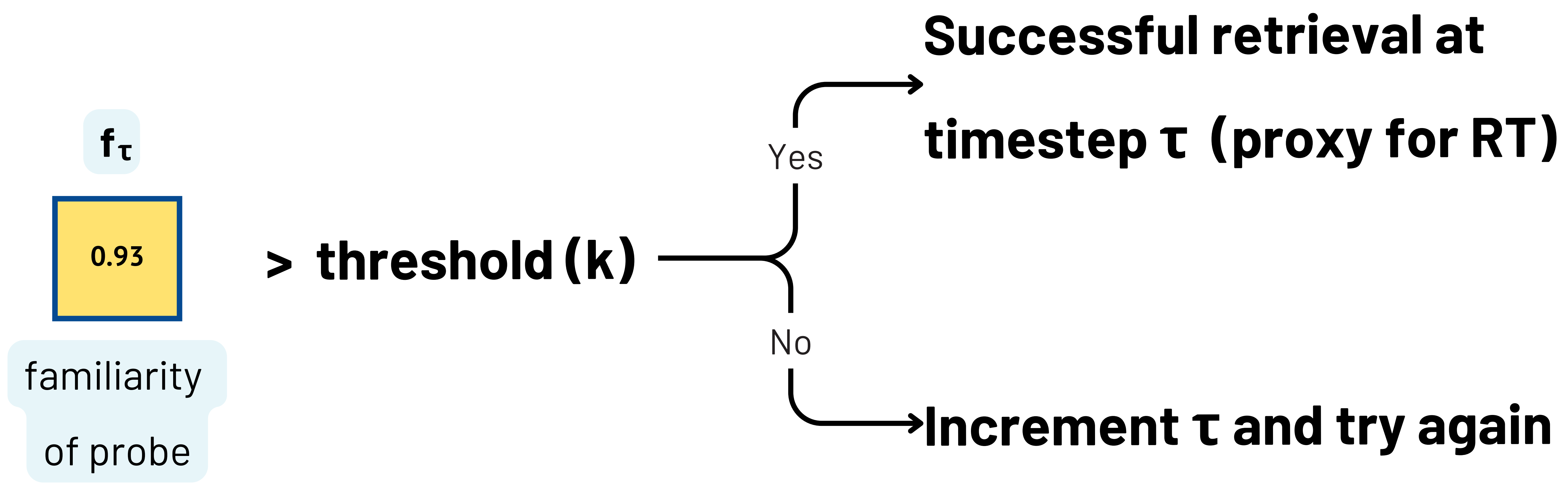


$$e_\tau = a_\tau \mathbf{M}$$

$$f_\tau = sim(e_\tau , p))$$

Get similarity of reconstruction to the probe with dot product





$f_\tau$
0.93
familiarity of probe
> threshold (k)
Yes
No
Successful retrieval at timestep τ (proxy for RT)
Increment τ and try again

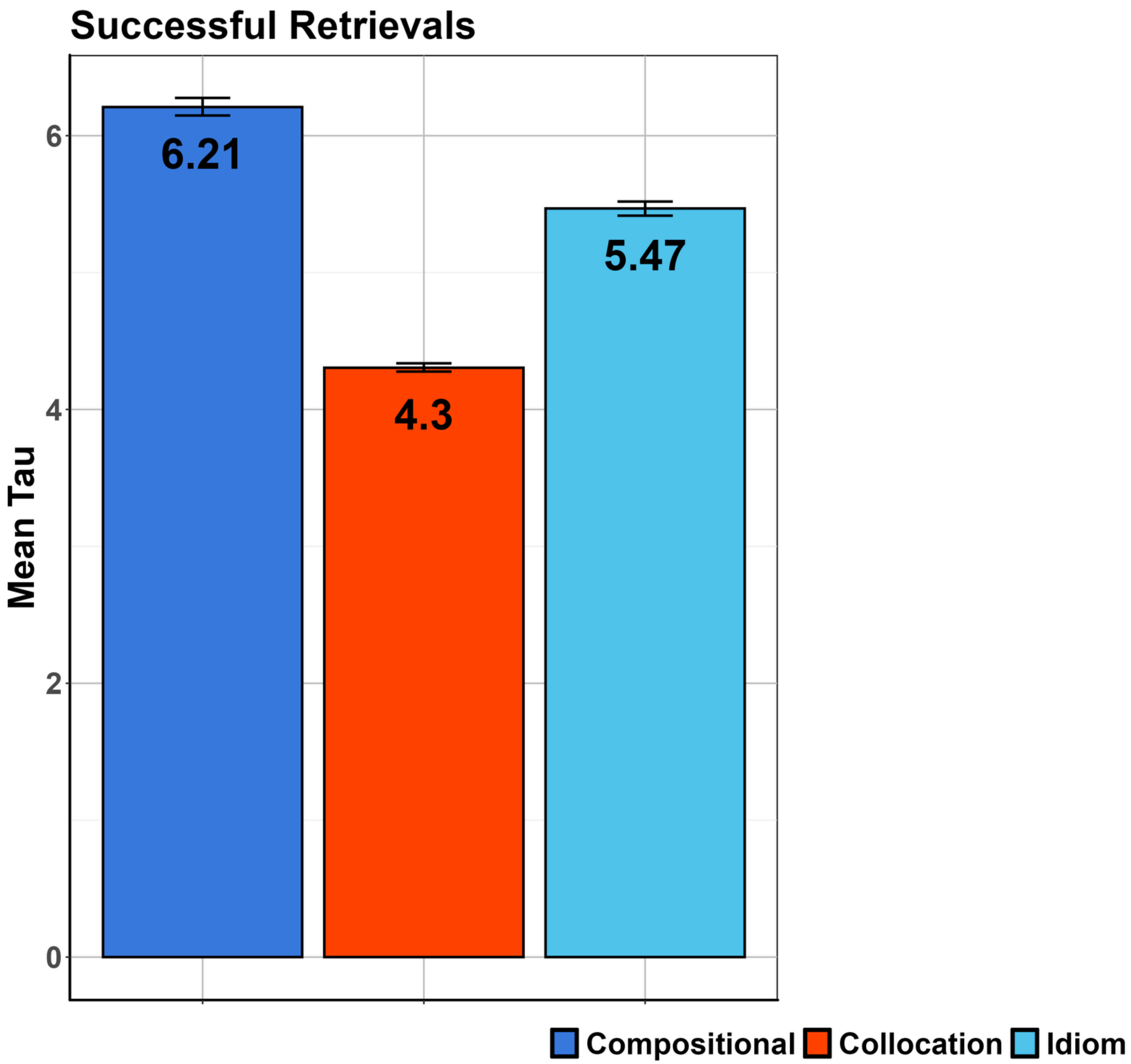
# SIMULATIONS

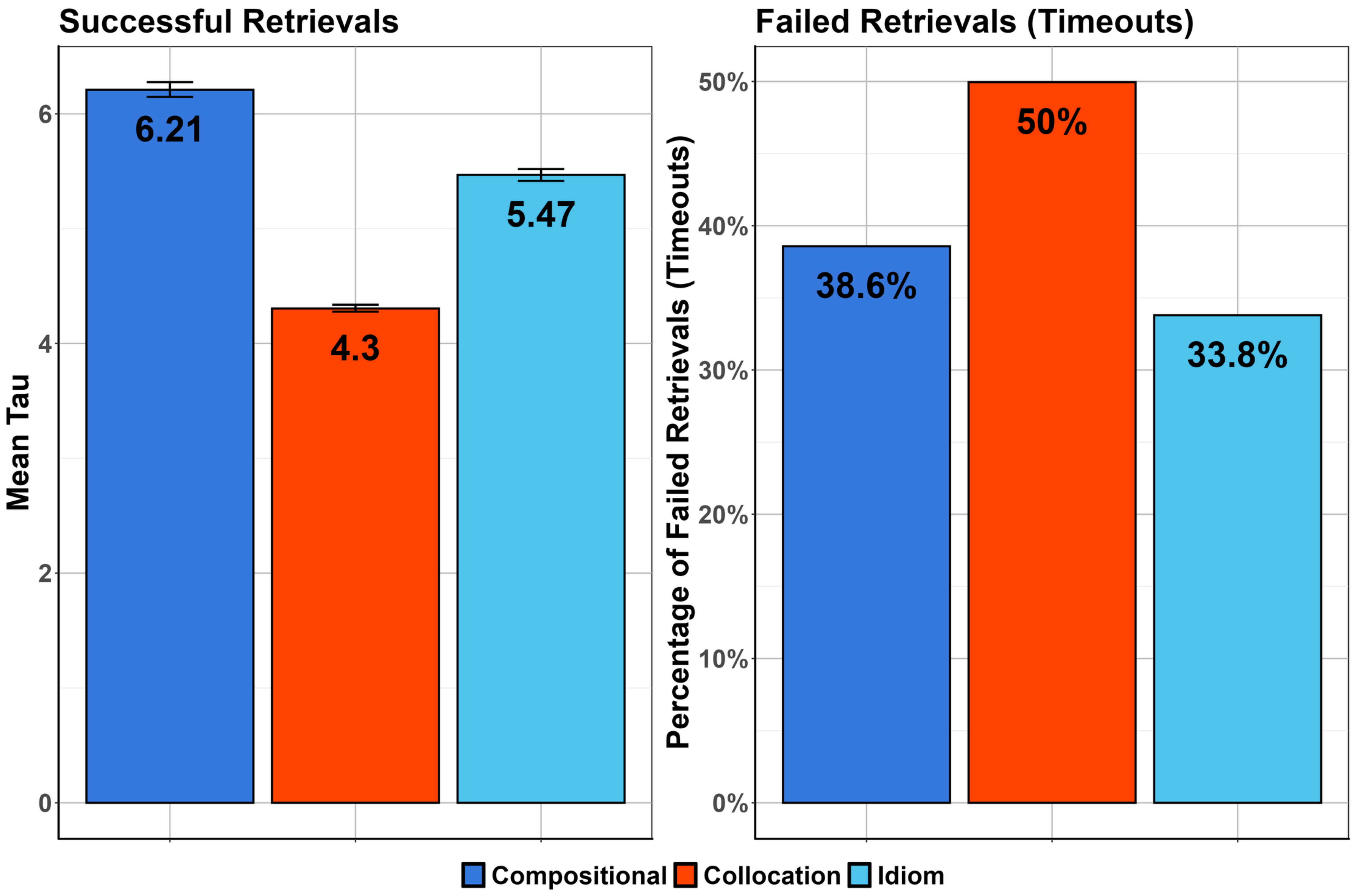
- **2 Hyperparameters:**
  - Retrieval threshold k (cosine sim > k = successful retrieval)
  - Forgetting probability (noise probability)
- **Attempt to retrieve each item**
  - Keep track of tau for successful retrieval
  - Timeout (max tau exceeded) = unsuccessful retrieval

# EXPERIMENTAL MANIPULATIONS

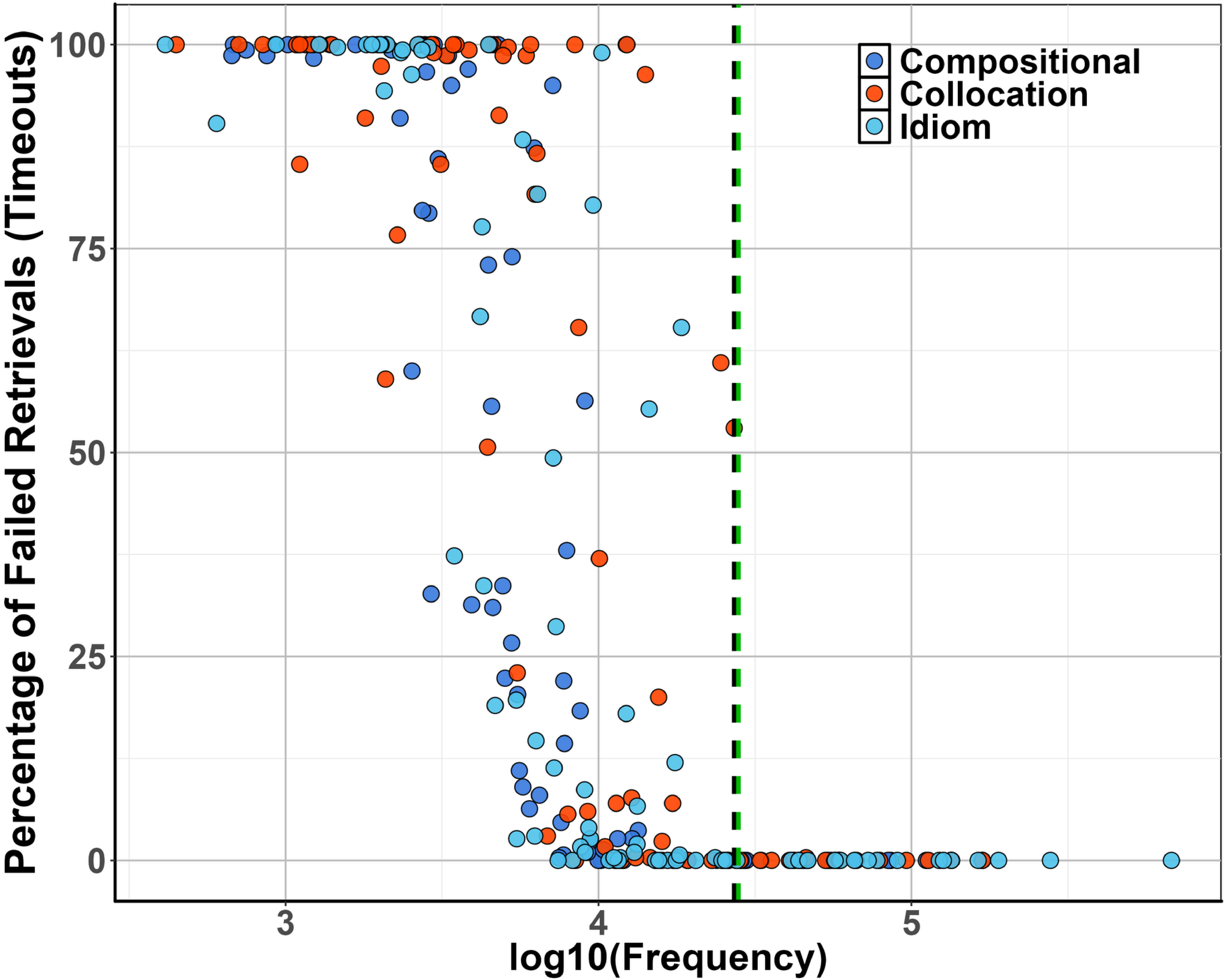
- **Main Experiment: Frequency & Semantics**
- **Sanity Checks:**
  - **Frequency-only**
    - Embedding vectors for each item replaced with unique random noise
  - **Semantics-only**
    - Each item is sampled into MINERVA memory in equal proportions
  - **Null model**
    - Equal frequency AND noise embeddings

# MODELLING RESULTS


Successful Retrievals
Mean Tau
6.21
4.3
5.47
0
2
4
6
Compositional
Collocation
Idiom



Successful Retrievals
Failed Retrievals (Timeouts)
Mean Tau
Percentage of Failed Retrievals (Timeouts)
6.21
4.3
5.47
38.6%
50%
33.8%
0
2
4
6
0%
10%
20%
30%
40%
50%
Compositional
Collocation
Idiom



Percentage of Failed Retrievals (Timeouts)
100
75
50
25
0
3
4
5
log10(Frequency)
Compositional
Collocation
Idiom

# SUMMARY

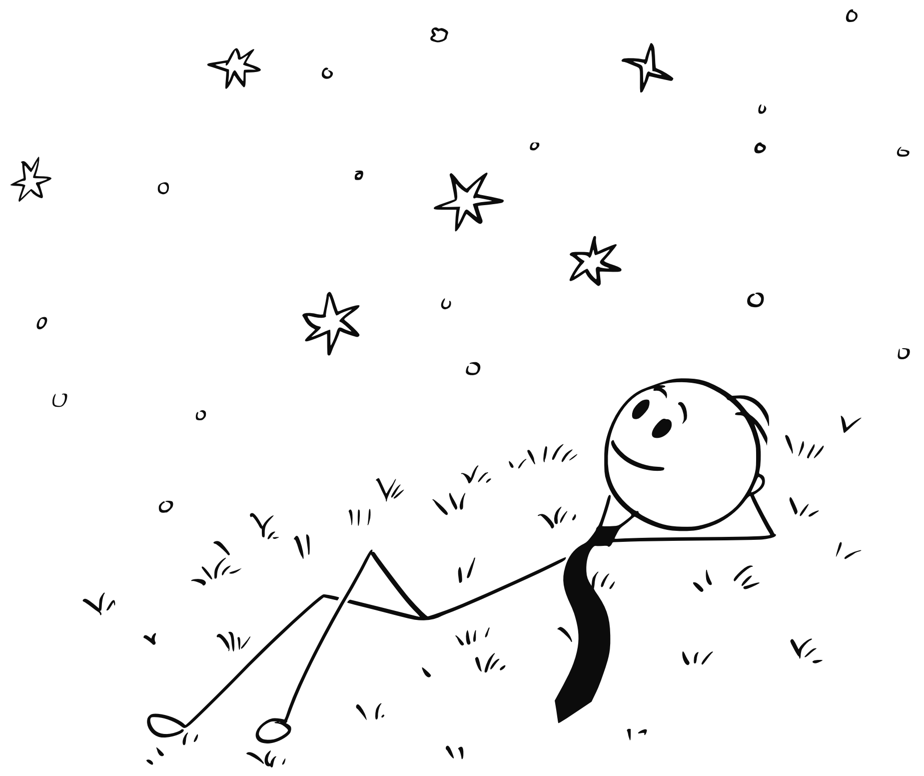
1. **Human results** show:
    a. Collocations are processed the slowest
    b. Marginal difference between idioms and productive items

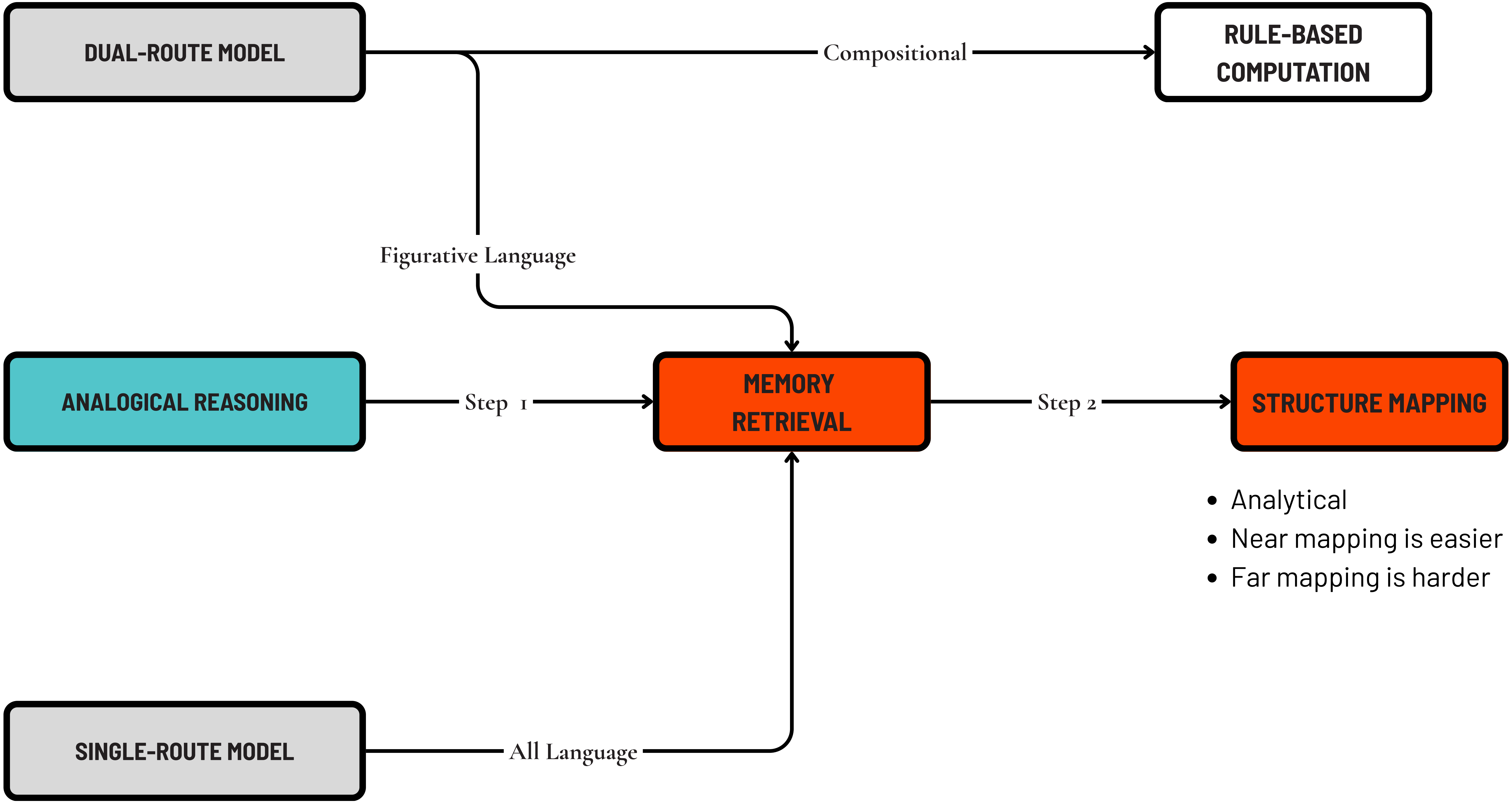
2. In **MINERVA2:**
    a. frequency threshold similar to humans
    b. above threshold, MINERVA mirrors human results
    c. below threshold, failures to retrieve mirror human results

# CONCLUSION

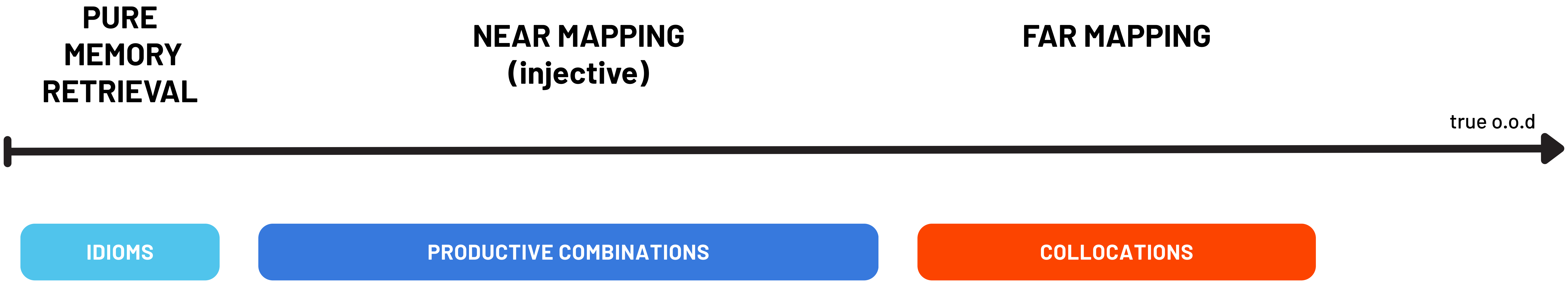
Memory retrieval is insufficient to capture human processing trends observed across the semantic compositionality continuum.

# SO, WHAT ELSE IS THERE?






Wray, A. (2002). Formulaic Language and the Lexicon. In Language (Vol. 80). Cambridge University Press.
https://doi.org/10.1353/lan.2004.0209




**This could account for the observed processing costs...**

# CAN WE MODEL THIS?

# Back to #1:
# Every modeling choice is a theoretical commitment…

# Slide Graveyard

# LIMITATIONS

**Task (for humans) and embeddings (for MINERVA) do not fully capture figurative vs compositional meanings for idioms:**

- Can we give humans a memory task which elicits figurative readings of idioms?
- Can we get embeddings that better capture figurativeness?