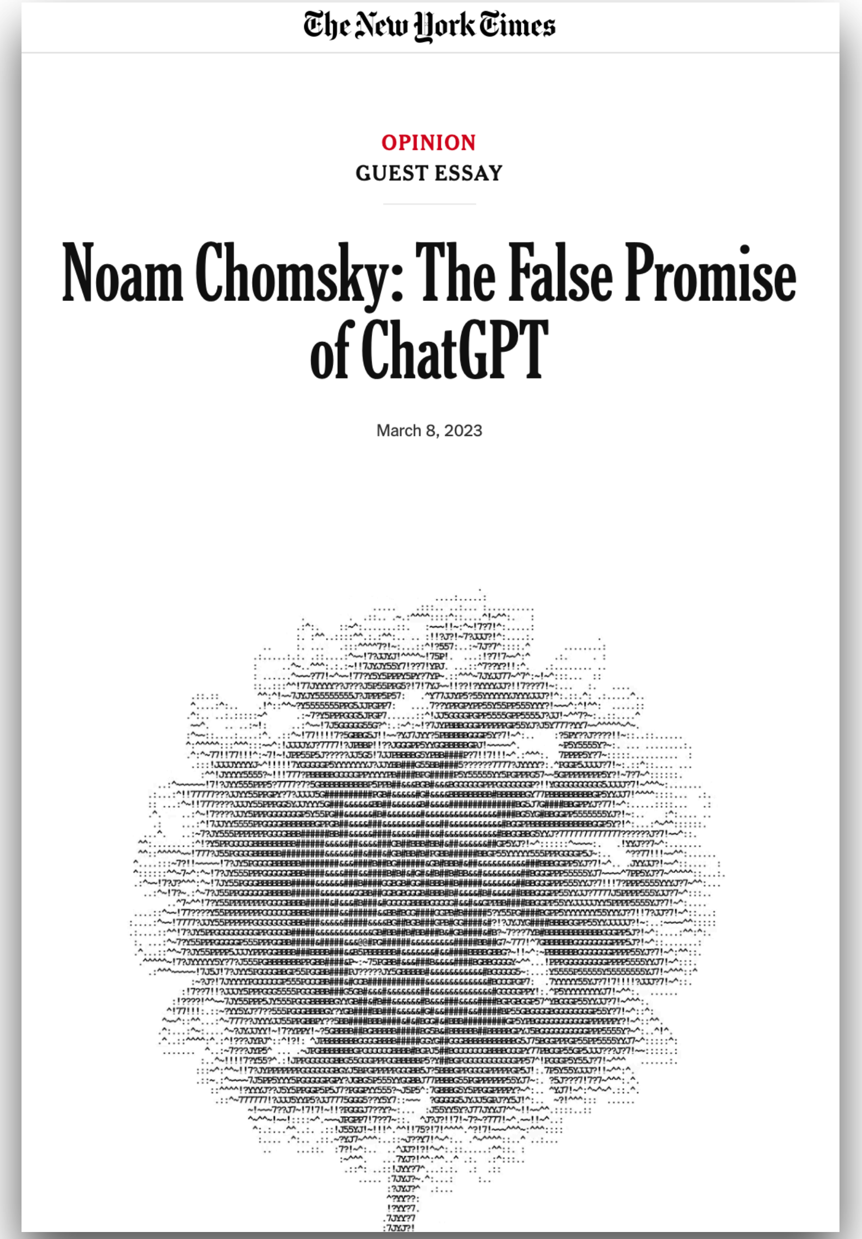
Generalised measures of predictive uncertainty in online language processing

Mario Giulianelli

Cambridge NLIP Seminar, 6 March 2026

What is the role for language models in linguistic theory?


The New York Times

OPINION

GUEST ESSAY

Noam Chomsky: The False Promise of ChatGPT

March 8, 2023

“Unlike humans, for example, who are endowed with a universal grammar that limits the languages we can learn to those with a certain kind of almost mathematical elegance, these programs **learn humanly possible and humanly impossible languages with equal facility**.

(Chomsky, Roberts, Watumull, 2023)

LM Skeptic

LMs won’t contribute to linguistic theory

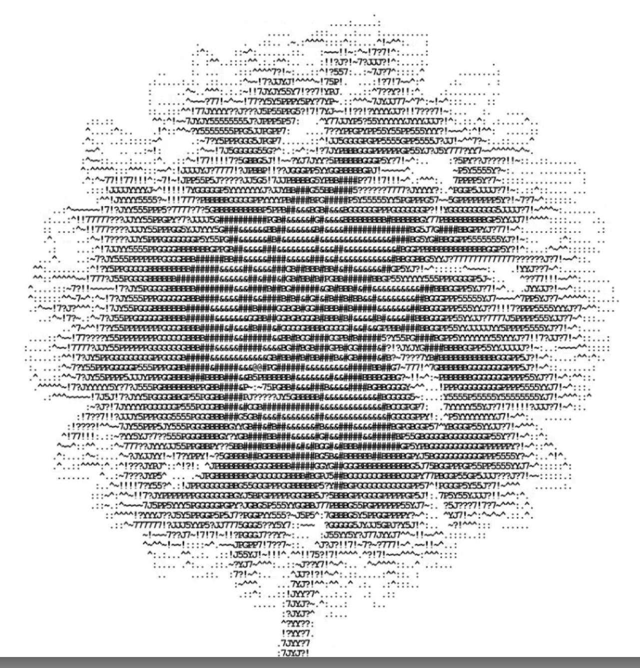
What is the role for language models in linguistic theory?

The New York Times

OPINION
GUEST ESSAY

Noam Chomsky: The False Promise of ChatGPT

March 8, 2023



"Unlike humans, for example, who are endowed with a universal grammar that limits the languages we can learn to those with a certain kind of almost mathematical elegance, these programs **learn humanly possible and humanly impossible languages with equal facility**.

(Chomsky, Roberts, Watumull, 2023)

"[…] their parameters come to **embody a theory of language**, including representations of latent state through a sentence and a discourse. The exact same logic of tuning parameters to formalize and then compare theories is found in other sciences, like modeling hurricanes or pandemics…"

(Piantadosi, 2023)

Modern language models refute Chomsky's approach to language

Steven T. Piantadosi
UC Berkeley & Helen Wills Neuroscience Institute

Modern machine learning has subverted and bypassed the theoretical framework of Chomsky's generative approach to linguistics, including its core claims to particular insights, principles, structures, and processes. I describe the sense in which modern language models implement genuine theories of language, and I highlight the links between these models and approaches to linguistics that are based on gradient computations and memorized constructions. I also describe why these models undermine strong claims for the innateness of language and respond to several critiques of large language models, including arguments that they can't answer "why" questions and skepticism that they are informative about real life acquisition. Most notably, large language models have attained remarkable success at discovering grammar without using any of the methods that some in linguistics insisted were necessary for a science of language to progress.

1 Introduction

After decades of privilege and prominence in linguistics, Noam Chomsky's approach to the science of language is experiencing a remarkable downfall. The story is, in part, a cautionary tale about what happens when an academic field isolates itself from what should be complementary endeavours. Chomsky's approach and methods are often argued to be problematic (e.g. Harris 1993, Pullum 1989, Behme 2012, Postal 2012, Behme 2014), but it is yet to be widely recognized just how the underlying ideas have been undermined by recent computational advances.

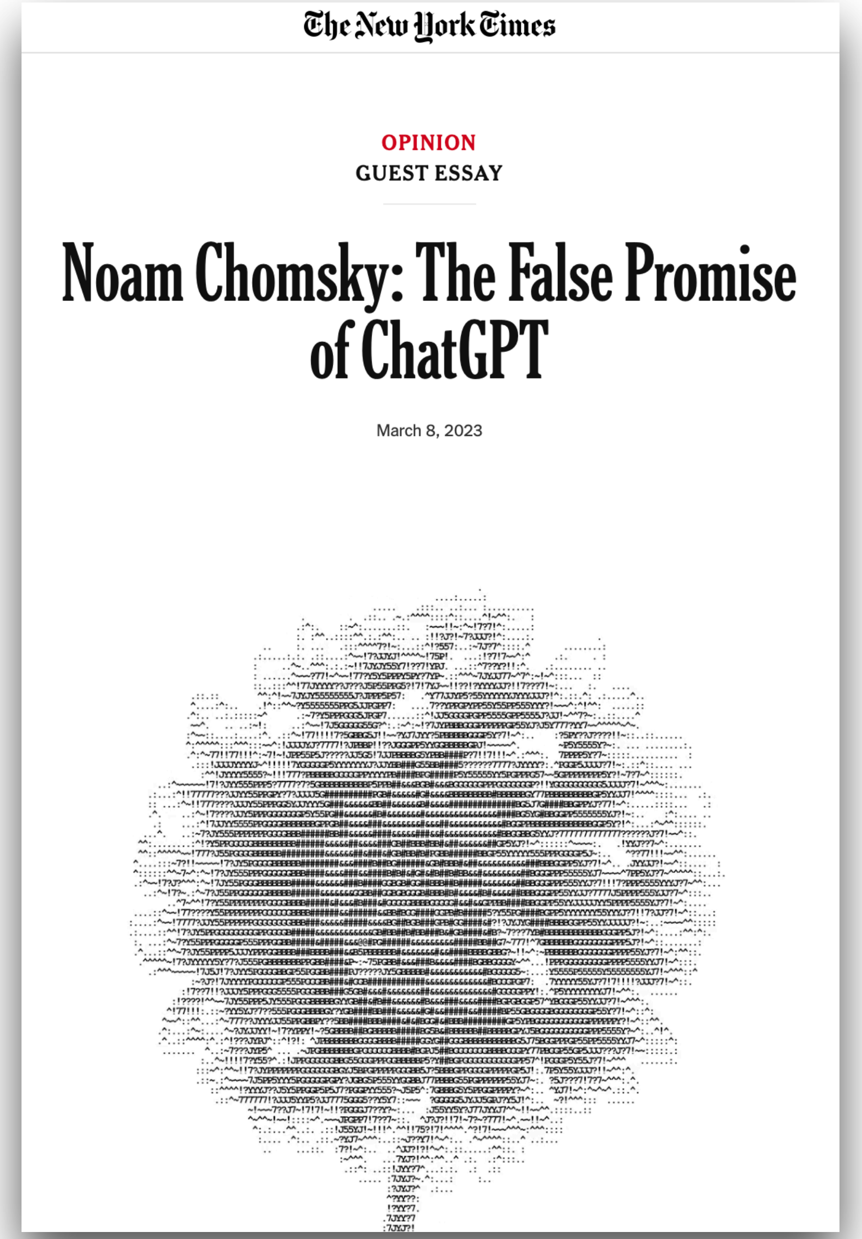
LM Skeptic

LM Enthusiast

LMs won't contribute to linguistic theory

LMs represent a theoretical paradigm shift

What is the role for language models in linguistic theory?


The New York Times

OPINION
GUEST ESSAY

Noam Chomsky: The False Promise of ChatGPT

March 8, 2023

"Unlike humans, for example, who are endowed with a universal grammar that limits the languages we can learn to those with a certain kind of almost mathematical elegance, these programs **learn humanly possible and humanly impossible languages with equal facility**."

(Chomsky, Roberts, Watumull, 2023)

"[…] their parameters come to **embody a theory of language**, including representations of latent state through a sentence and a discourse. The exact same logic of tuning parameters to formalize and then compare theories is found in other sciences, like modeling hurricanes or pandemics…"

(Piantadosi, 2023)

Modern language models refute Chomsky's approach to language

Steven T. Piantadosi

UC Berkeley & Helen Wills Neuroscience Institute

Modern machine learning has subverted and bypassed the theoretical framework of Chomsky's generative approach to linguistics, including its core claims to particular insights, principles, structures, and processes. I describe the sense in which modern language models implement genuine theories of language, and I highlight the links between these models and approaches to linguistics that are based on gradient computations and memorized constructions. I also describe why these models undermine strong claims for the innateness of language and respond to several critiques of large language models, including arguments that they can't answer "why" questions and skepticism that they are informative about real life acquisition. Most notably, large language models have attained remarkable success at discovering grammar without using any of the methods that some in linguistics insisted were necessary for a science of language to progress.

1 Introduction

After decades of privilege and prominence in linguistics, Noam Chomsky's approach to the science of language is experiencing a remarkable downfall. The story is, in part, a cautionary tale about what happens when an academic field isolates itself from what should be complementary endeavours. Chomsky's approach and methods are often argued to be problematic (e.g. Harris 1993, Pullum 1989, Behme 2012, Postal 2012, Behme 2014), but it is yet to be widely recognized just how the underlying ideas have been undermined by recent computational advances.

LM Skeptic

LM Optimist

LM Enthusiast

LMs won't contribute to linguistic theory

Language models can be used to empirically test and refine theories of language.

LMs represent a theoretical paradigm shift

What is the role for language models in linguistic theory?

Implementational

Kriegeskorte et al. 2008;
Marblestone et al. 2016;
Schrimpf et al. 2021

Algorithmic

Linzen et al. 2016; Gulordava et al. 2018;
Giulianelli et al. 2018; Hanna & Mueller 2024

Computational

Zipf 1965; Hale 2001; Levy 2008;
Gibson et al. 2019; Futrell & Hahn 2022

Observational

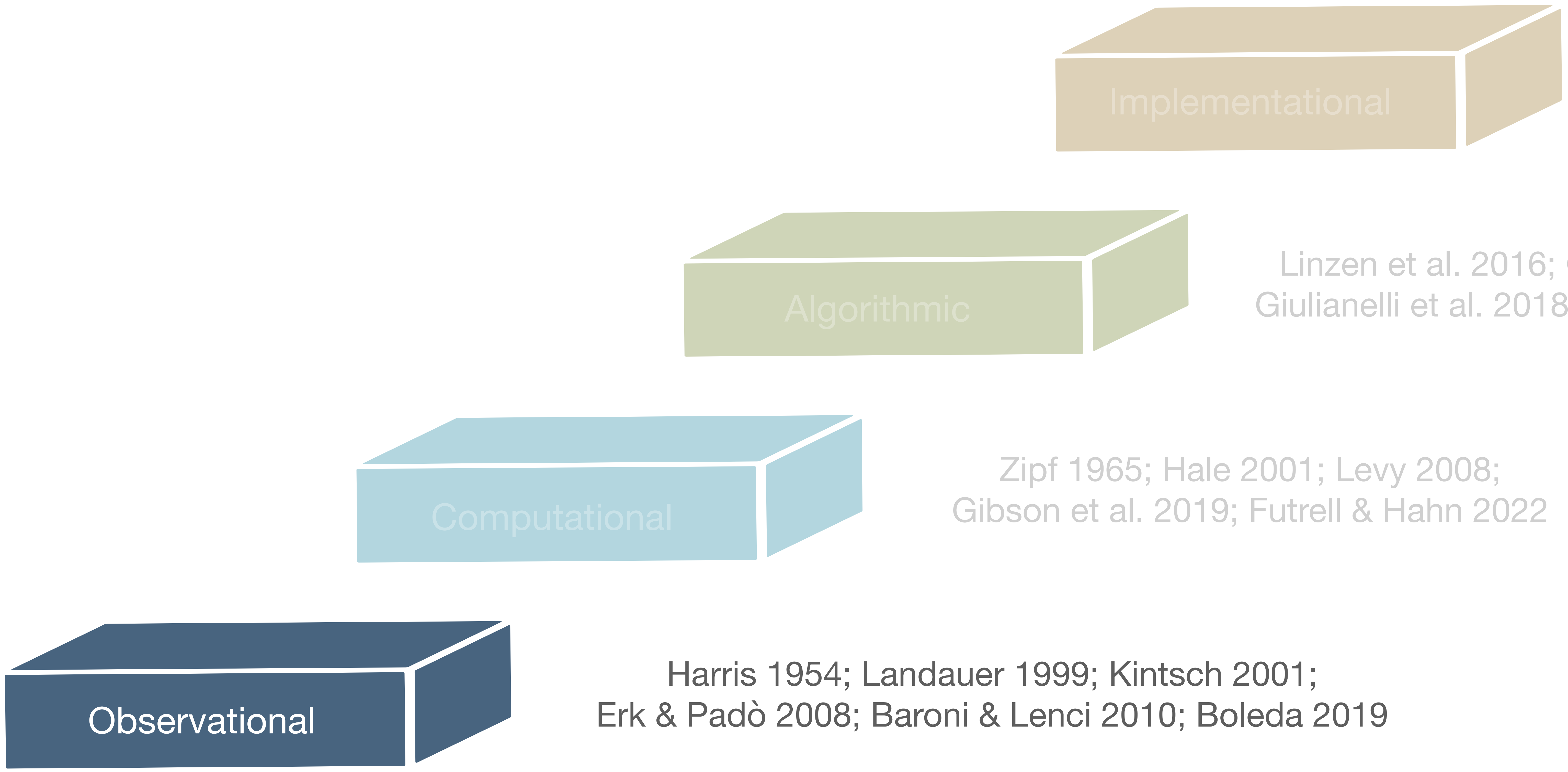
Harris 1954; Landauer 1999; Kintsch 2001;
Erk & Padò 2008; Baroni & Lenci 2010; Boleda 2019

LMs won’t contribute to linguistic theory

Language models can be used to empirically test and refine theories of language.

LMs represent a theoretical paradigm shift

What is the role for language models in linguistic theory?



Kriegeskorte et al. 2008; Marblestone et al. 2016; Schrimpf et al. 2021

Linzen et al. 2016; Gulordava et al. 2018; Giulianelli et al. 2018; Hanna & Mueller 2024

Zipf 1965; Hale 2001; Levy 2008; Gibson et al. 2019; Futrell & Hahn 2022

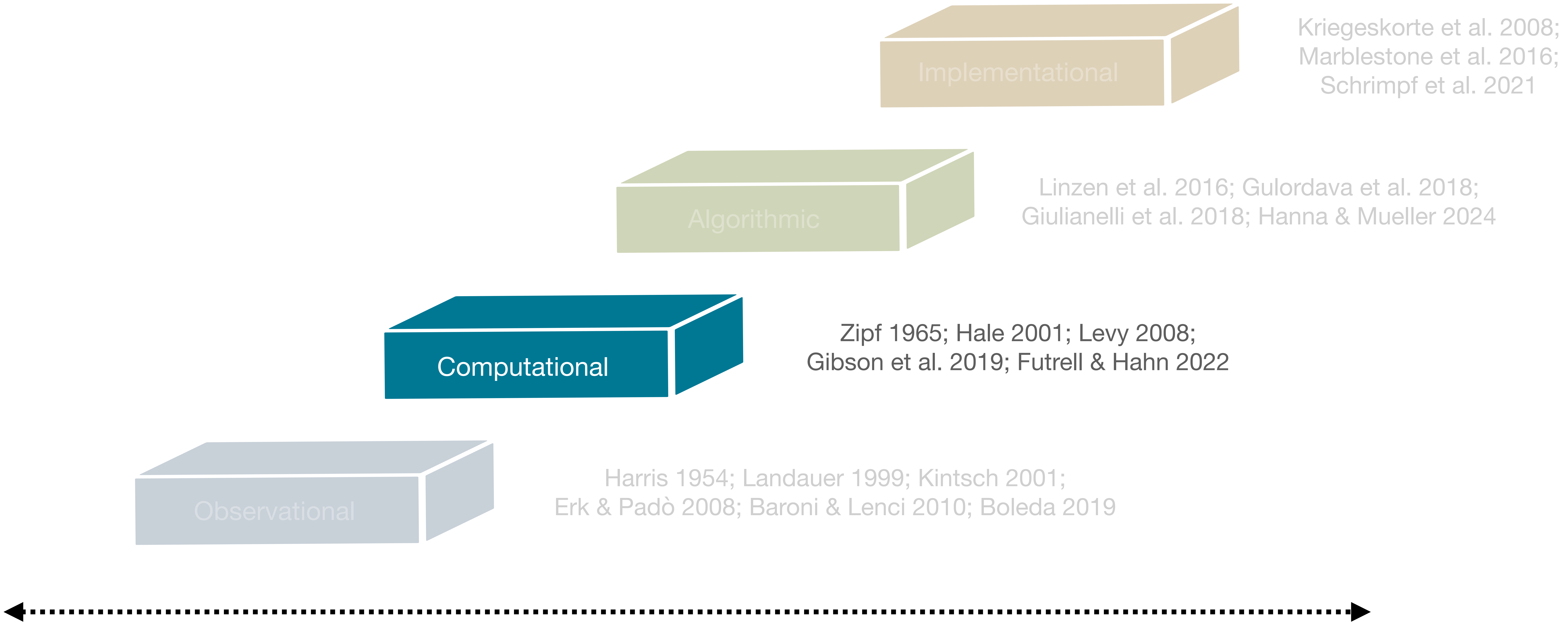
Harris 1954; Landauer 1999; Kintsch 2001; Erk & Padò 2008; Baroni & Lenci 2010; Boleda 2019

LMs won’t contribute to linguistic theory

Language models can be used to empirically test and refine theories of language.

LMs represent a theoretical paradigm shift

What is the role for language models in linguistic theory?

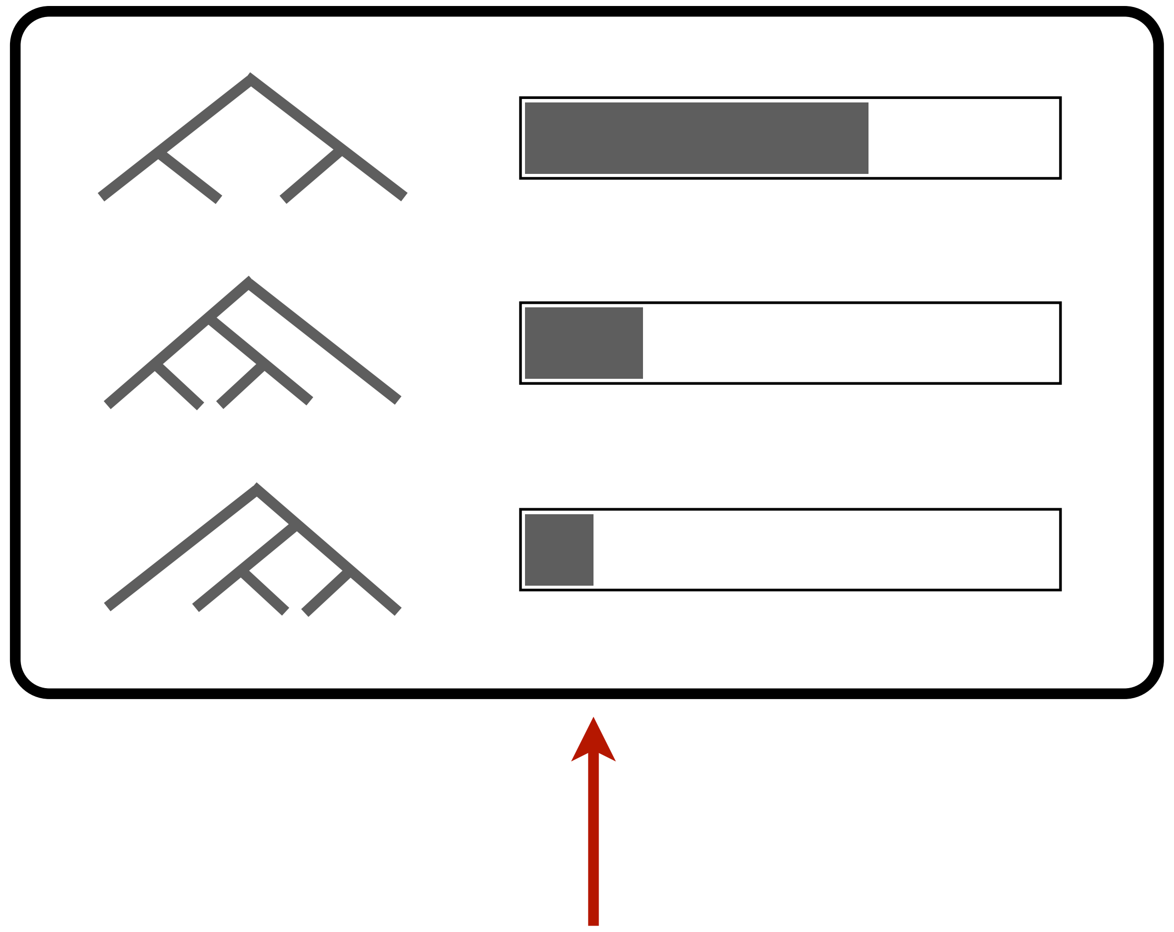


LMs won't contribute to linguistic theory

Language models can be used to empirically test and refine theories of language.

LMs represent a theoretical paradigm shift

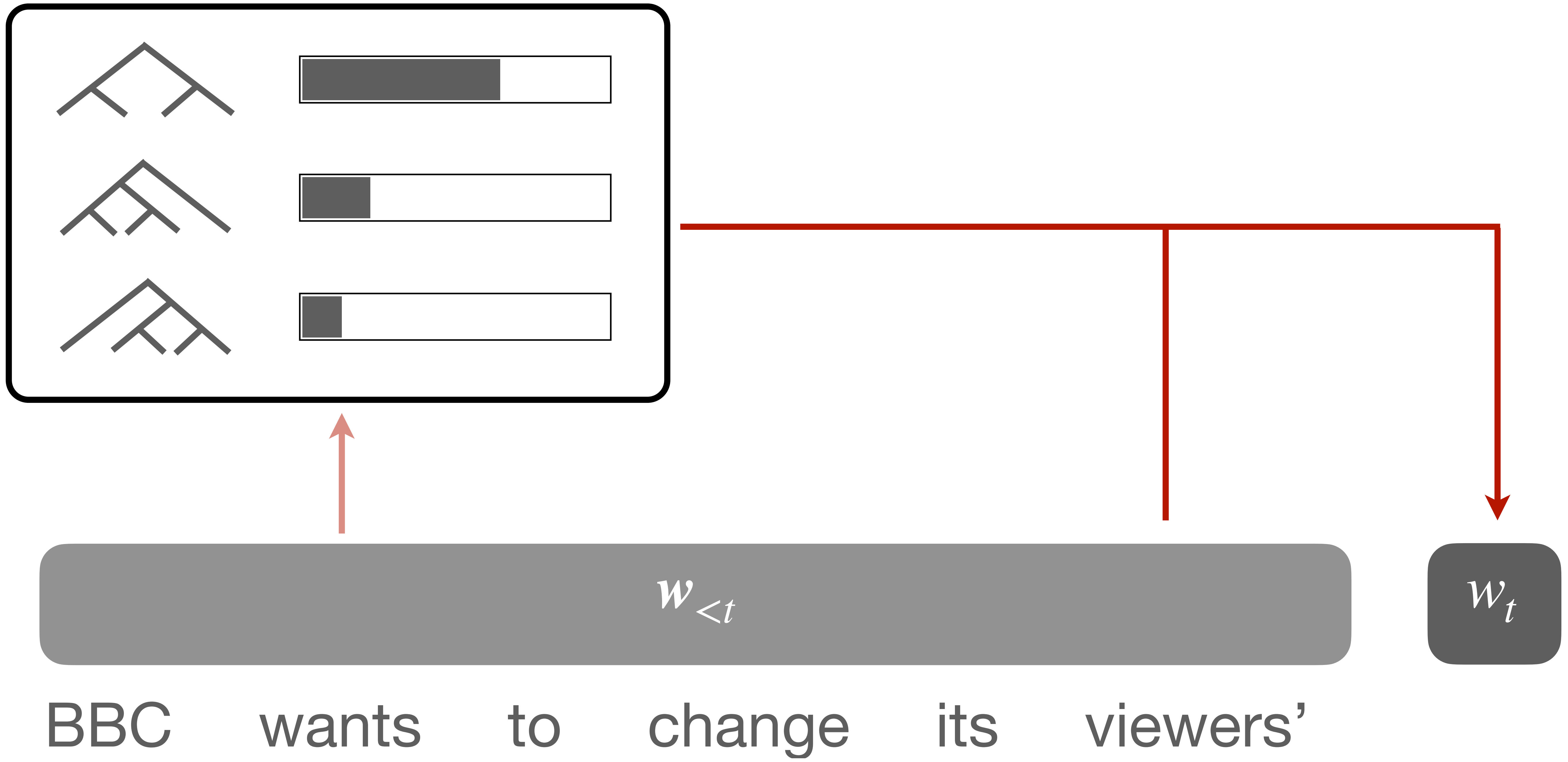
A (computational-level) story of incremental comprehension



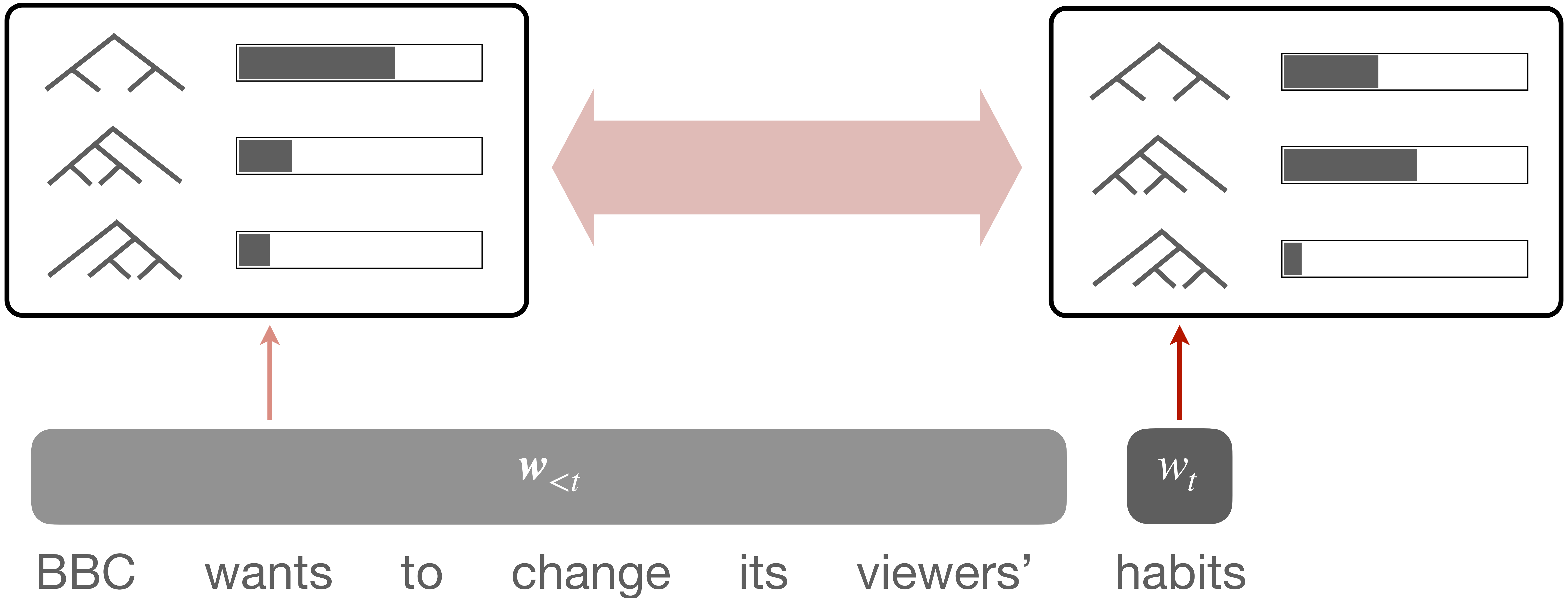
$w_{<t}$ w_t

BBC wants to change its viewers'

A (computational-level) story of incremental comprehension



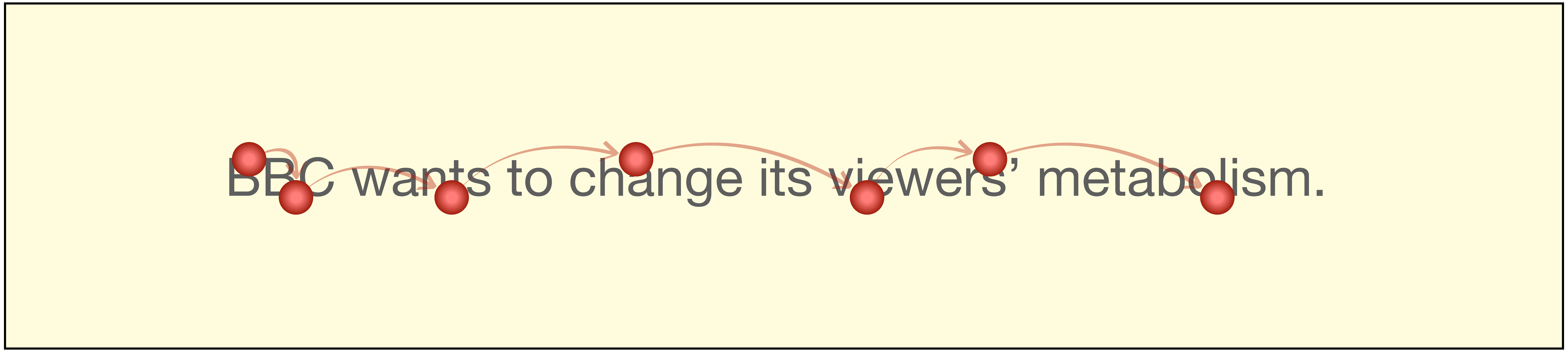
A (computational-level) story of incremental comprehension



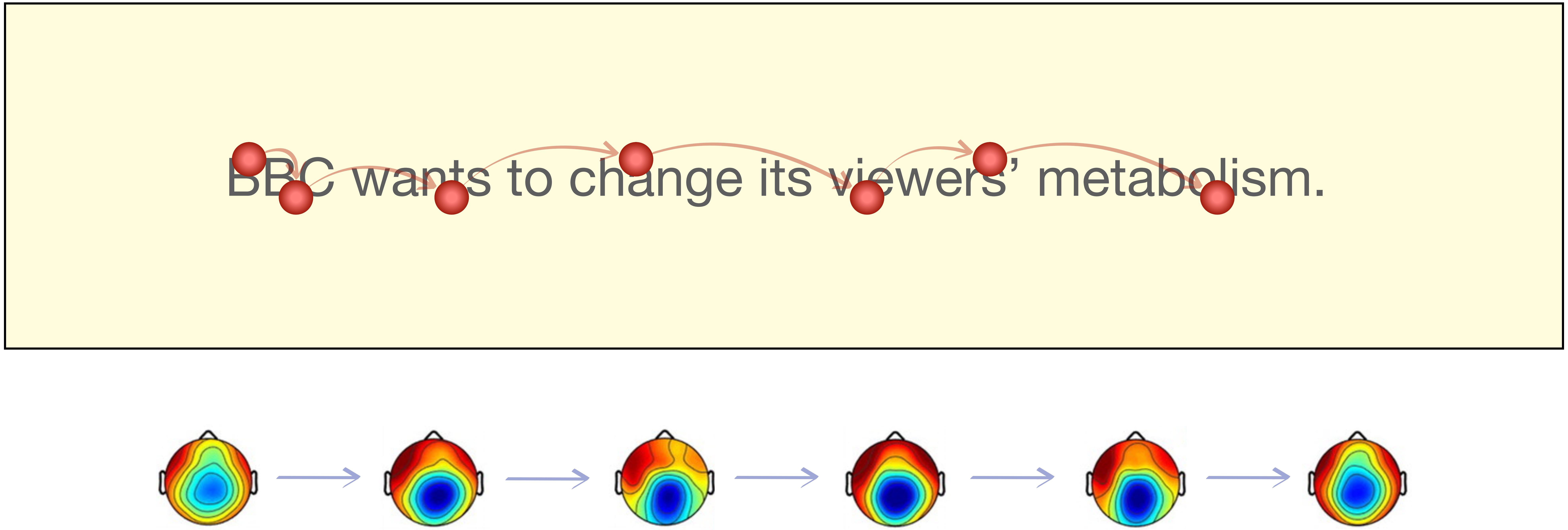
Incremental comprehension through reading experiments

BBC wants to change its viewers' metabolism.

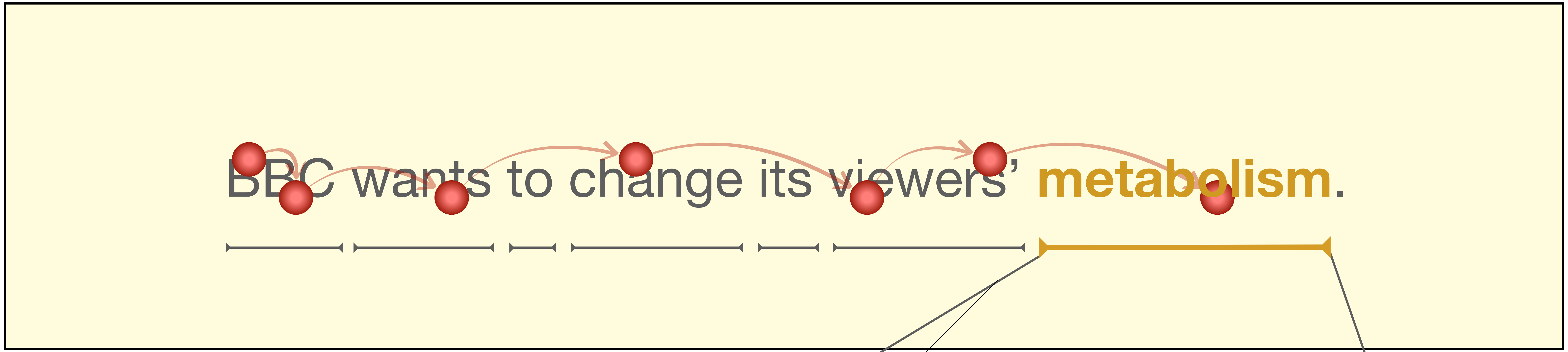
Incremental comprehension through reading experiments

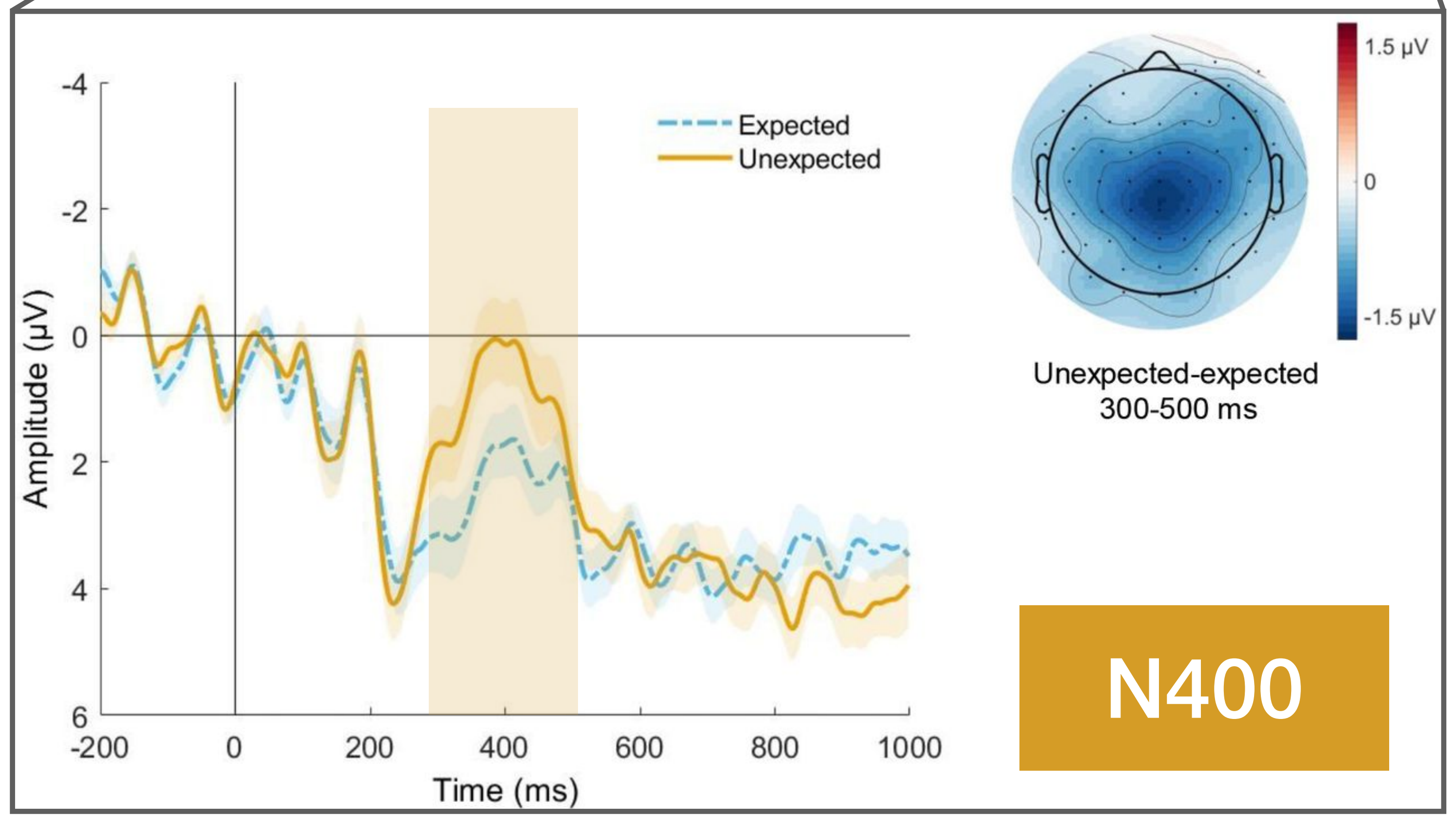


Incremental comprehension through reading experiments



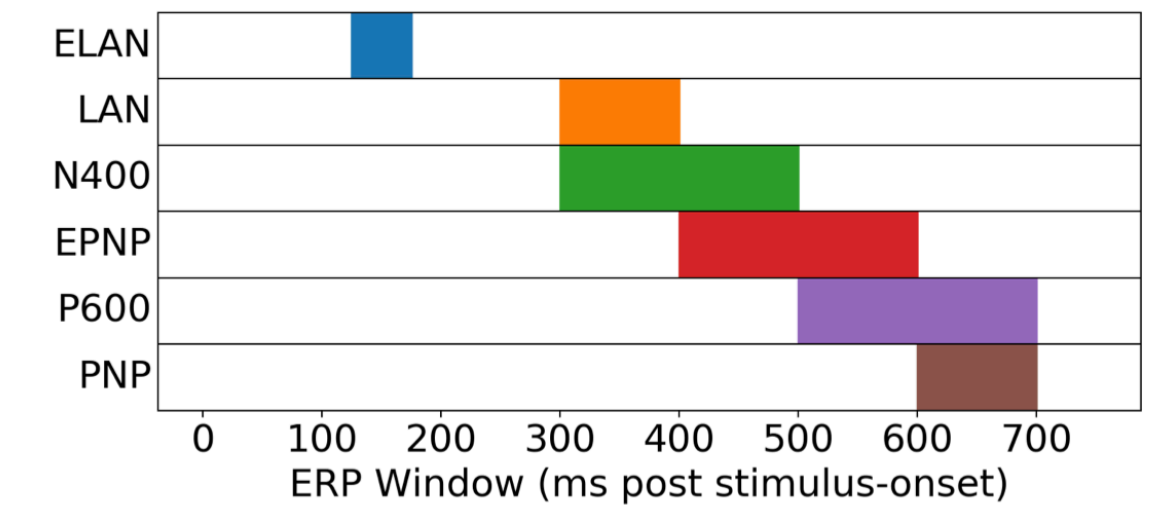
Behavioural and neural responses to linguistic input provide a direct window into the **cognitive processes underlying language comprehension**.

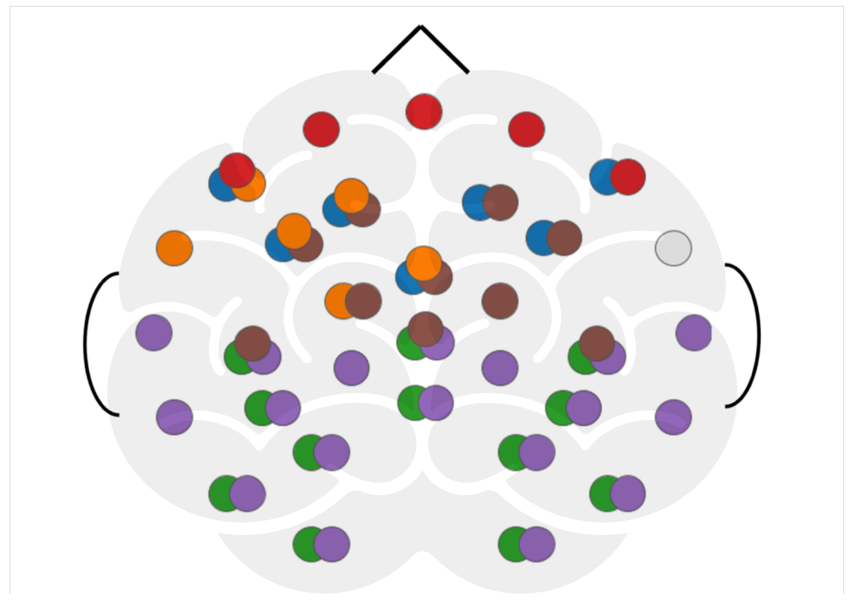




Incremental comprehension data

Event-related Brain Potentials (ERPs)





Incremental Stimuli

$\boldsymbol{w}_{<t}$						w_t
BBC	wants	to	change	its	viewers'	metabolism

Stimulus

$\boldsymbol{w}_{<t}$ w_t

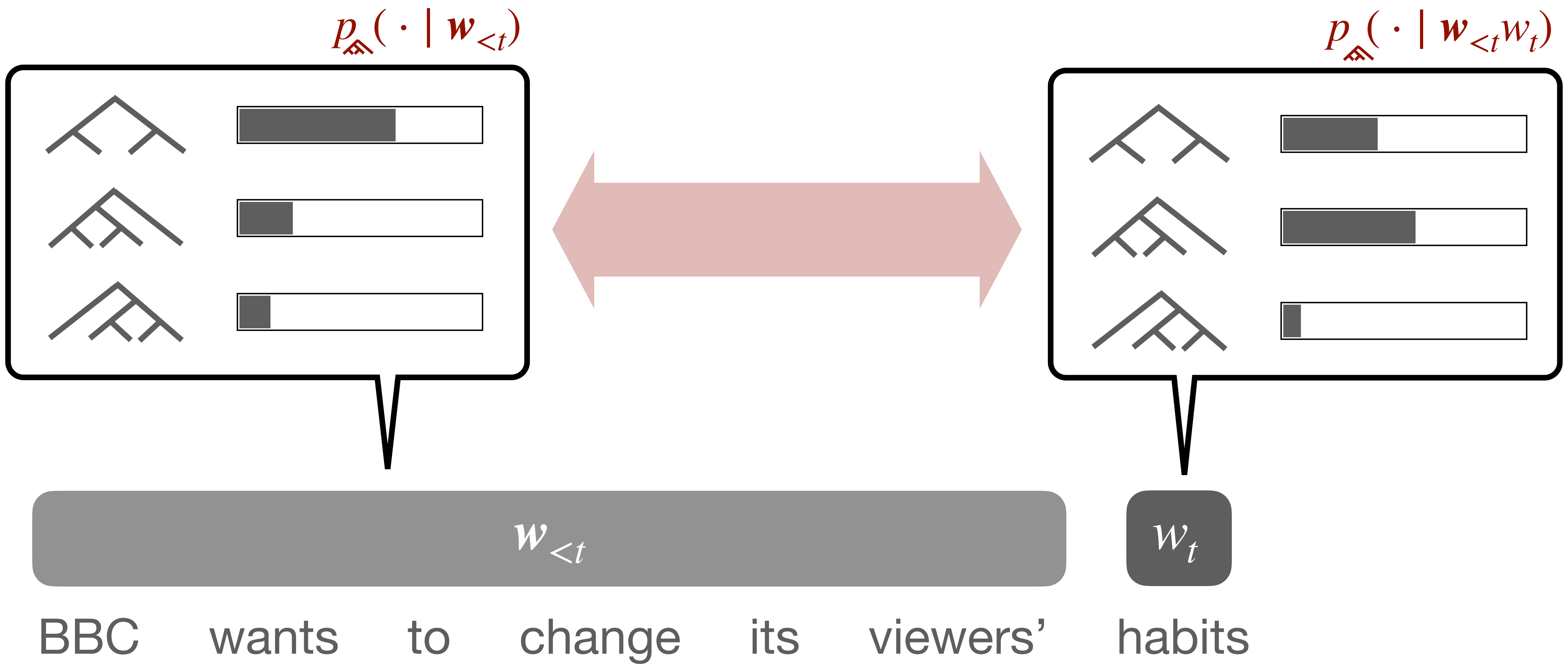
$t \in 1..T$

Response

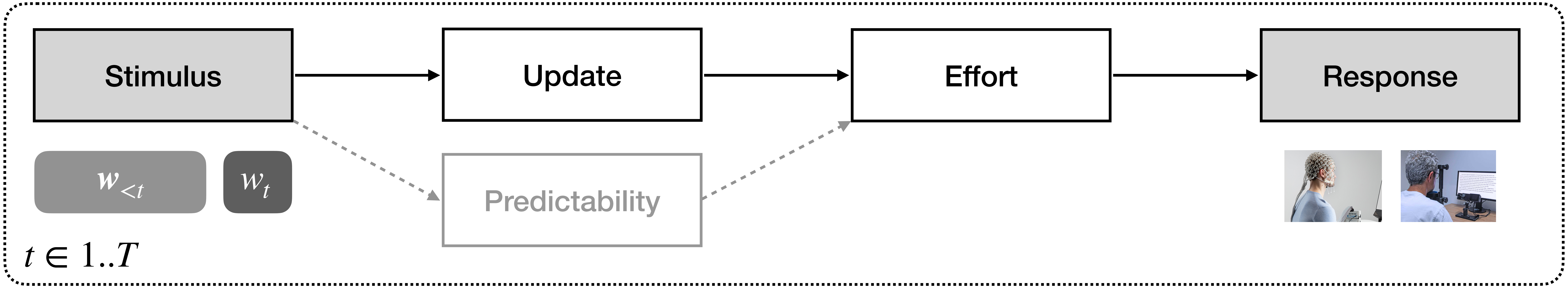
Brain imaging

Eye tracking

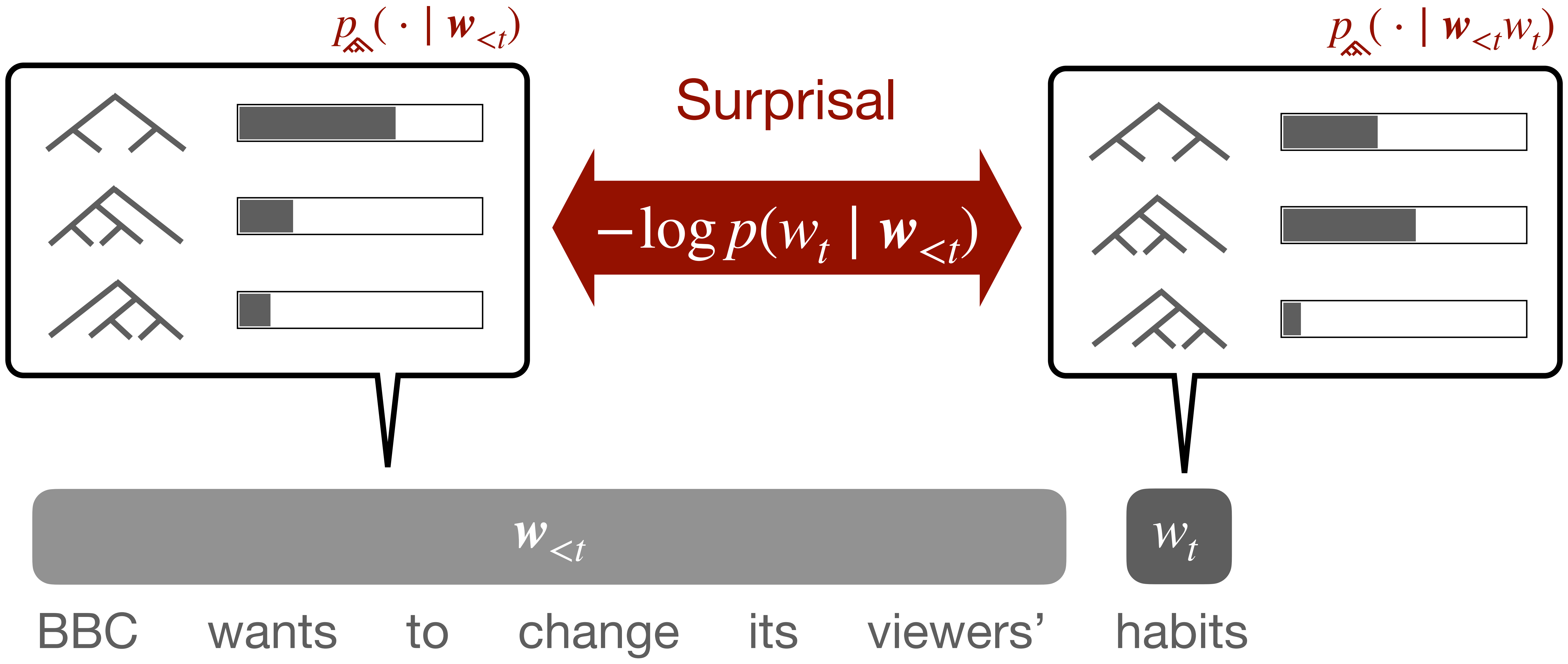
The Surprisal model of incremental comprehension



Response ∝ Effort ∝ Update



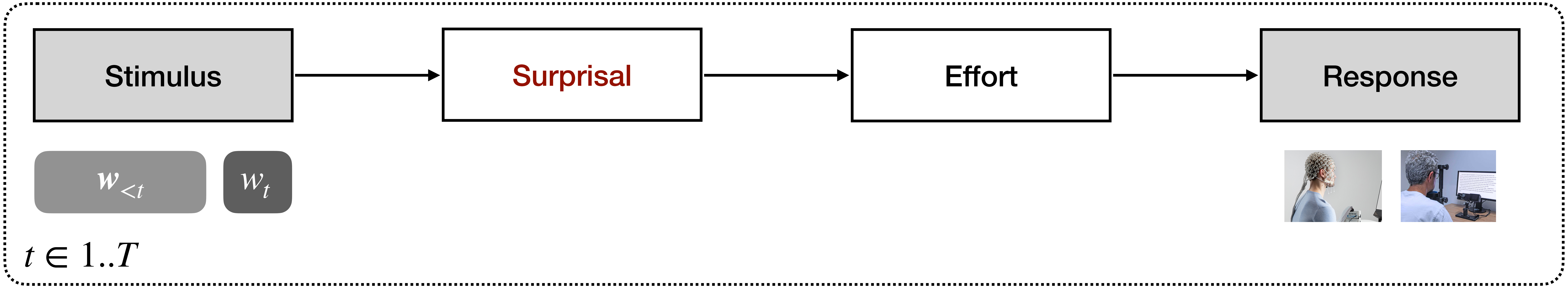
The Surprisal model of incremental comprehension



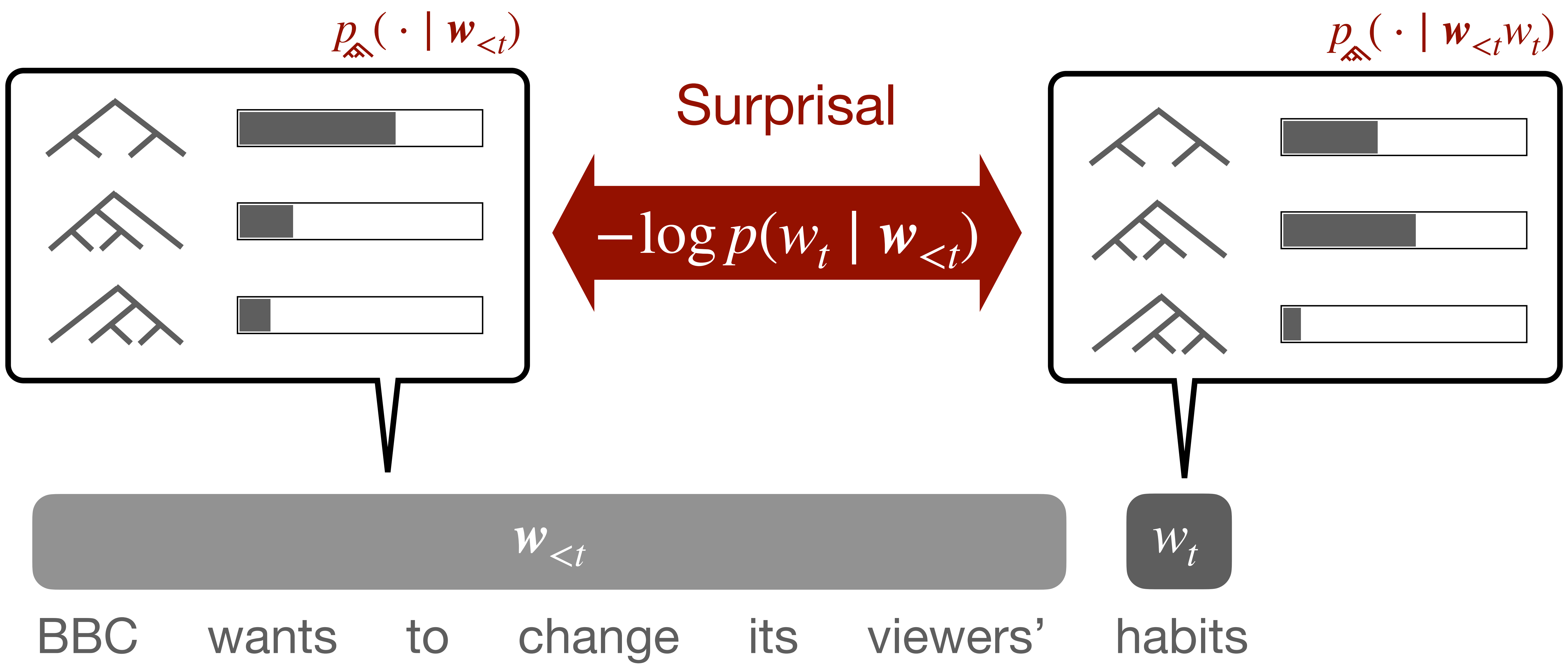
Response $\propto$ *Effort* $\propto$ *Update*

$\approx$ *Surprisal* $:= \boxed{-\log p(w_t \mid \boldsymbol{w}_{<t})}$

"human language model"



The Surprisal model of incremental comprehension

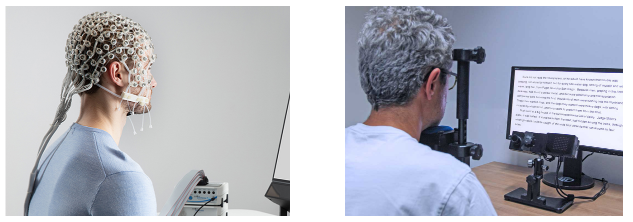


Response $\propto$ *Effort* $\propto$ *Update*

$\approx$ *Surprisal* $:= \boxed{-\log p_{LM}(w_t \mid \boldsymbol{w}_{<t})}$

parametrised language model

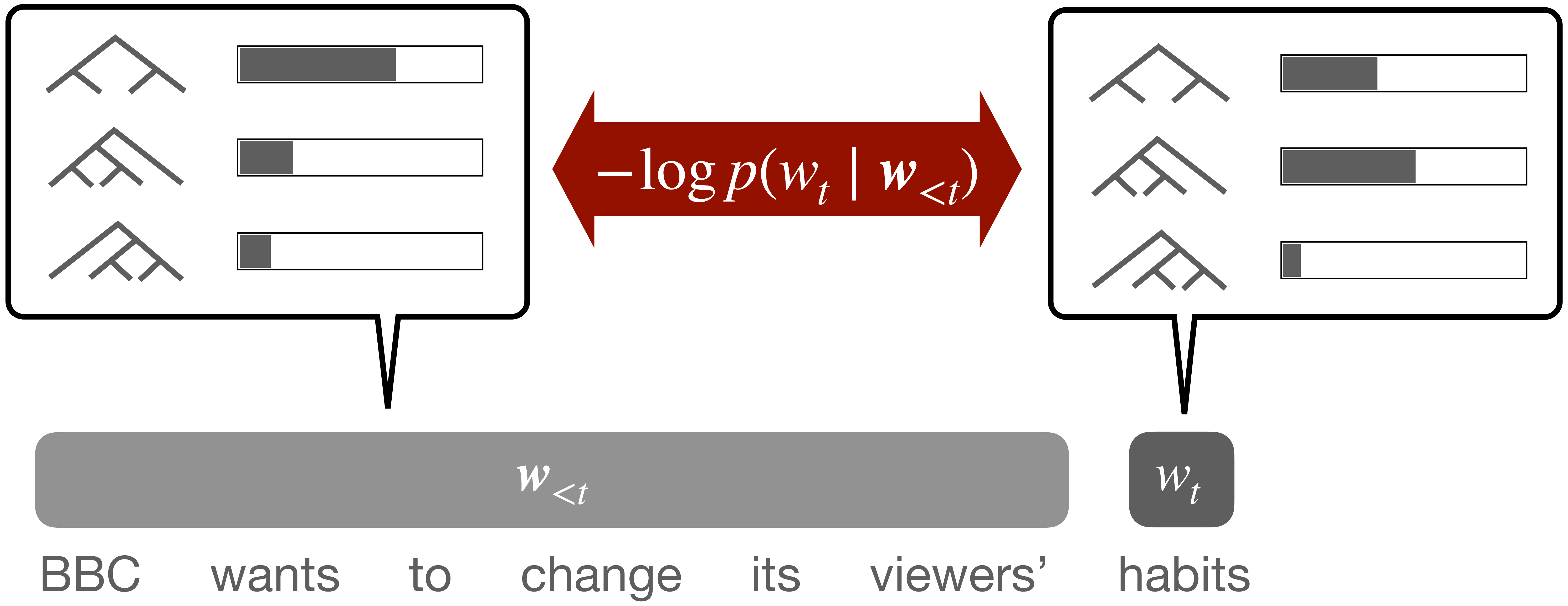
Stimulus → **Surprisal** → Effort → Response

$\boldsymbol{w}_{<t}$ w_t - - - - →  - - - - → 

Linear regression | GAM | Mixed-effect model

$t \in 1..T$

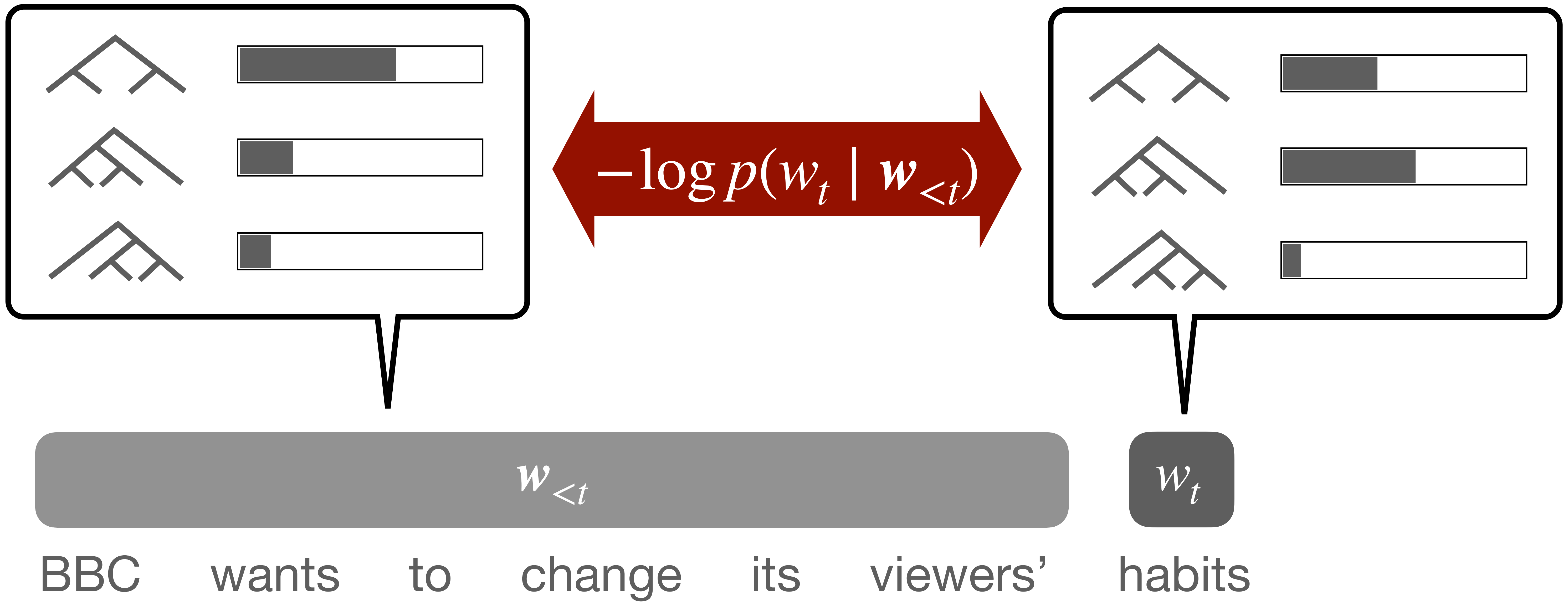
Destructuring Surprisal theory



Giulianelli, Wallbridge, Fernández. EMNLP 2023.

$$\iota^{\text{Surprisal}}(w_t; \boldsymbol{w}_{<t}) := -\log p(w_t \mid \boldsymbol{w}_{<t})$$

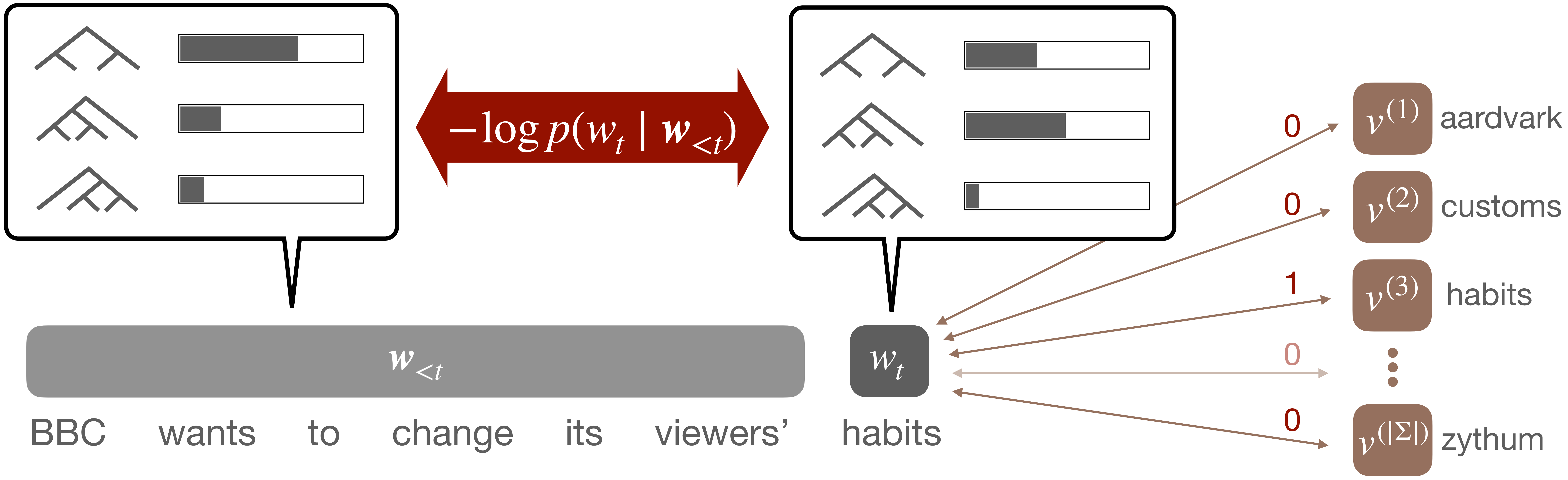
Destructuring Surprisal theory



Giulianelli, Wallbridge, Fernández. EMNLP 2023.

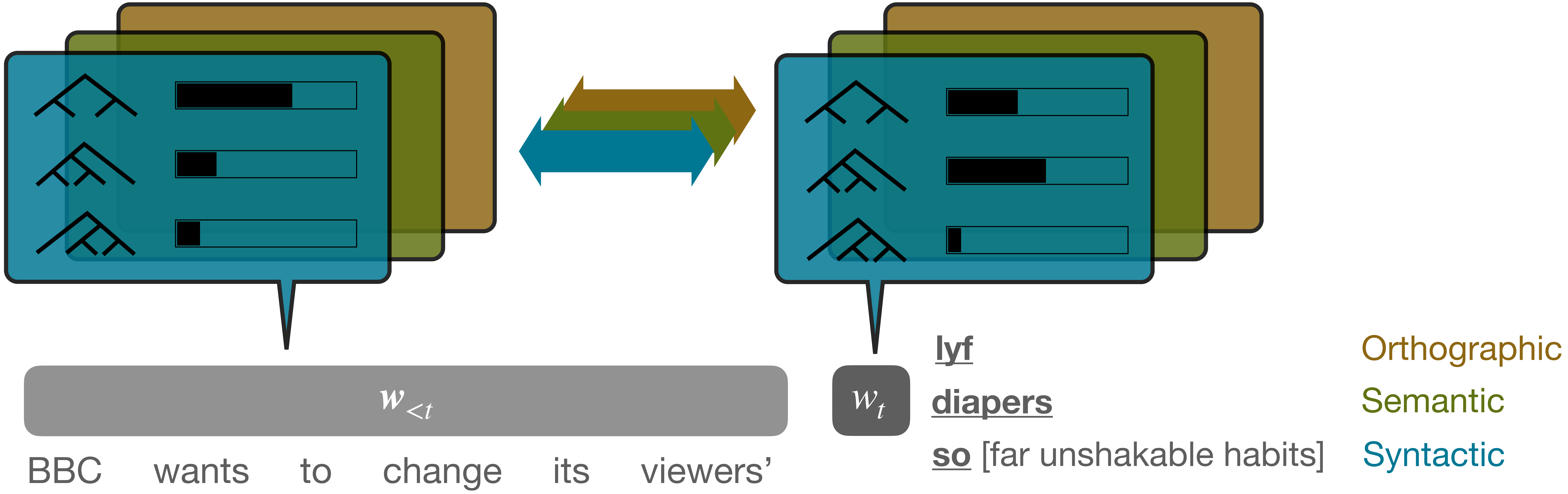
$$\iota^{\text{Surprisal}}(w_t; \boldsymbol{w}_{<t}) := -\log p(w_t \mid \boldsymbol{w}_{<t}) = -\log \sum_{v \in \Sigma} p(v \mid \boldsymbol{w}_{<t}) \boxed{\mathbf{1}\{v = w_t\}}$$

Destructuring Surprisal theory



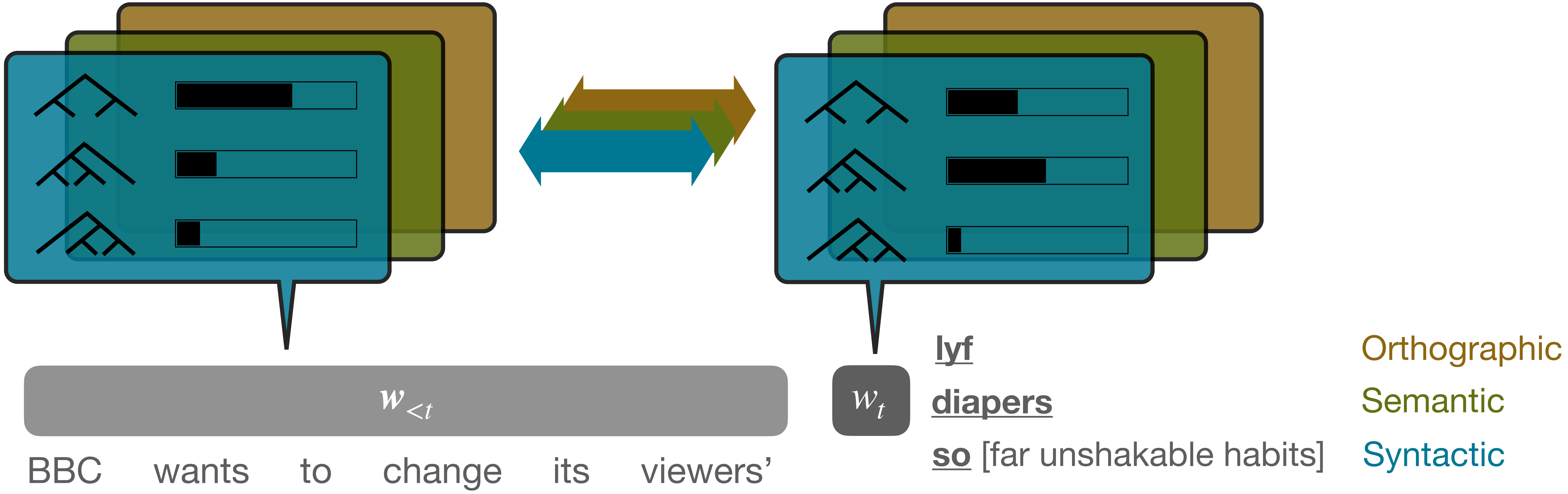
$$\iota^{\text{Surprisal}}(w_t; \boldsymbol{w}_{<t}) := -\log p(w_t \mid \boldsymbol{w}_{<t}) = -\log \boxed{\sum_{v \in \Sigma}} p(v \mid \boldsymbol{w}_{<t}) \boxed{\mathbf{1}\{v = w_t\}}$$

Destructuring Surprisal theory



$$\iota^{\text{Surprisal}}(w_t; \boldsymbol{w}_{<t}) := -\log p(w_t \mid \boldsymbol{w}_{<t}) = -\log \sum_{v \in \Sigma} p(v \mid \boldsymbol{w}_{<t}) \boxed{\mathbf{1}\{v = w_t\}}$$

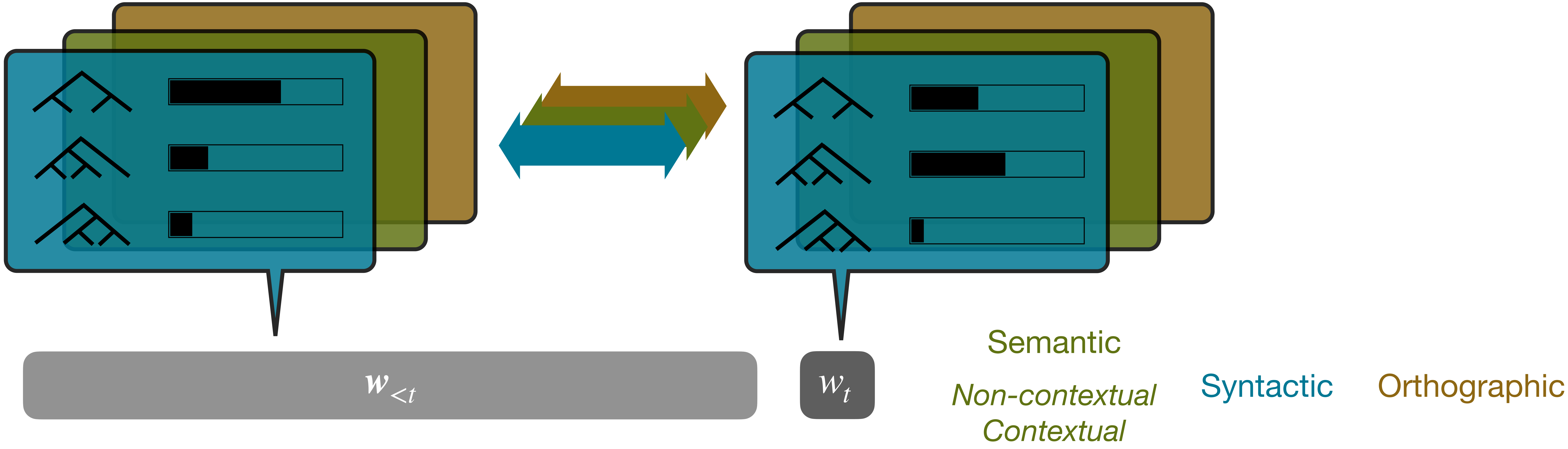
Similarity-adjusted Surprisal

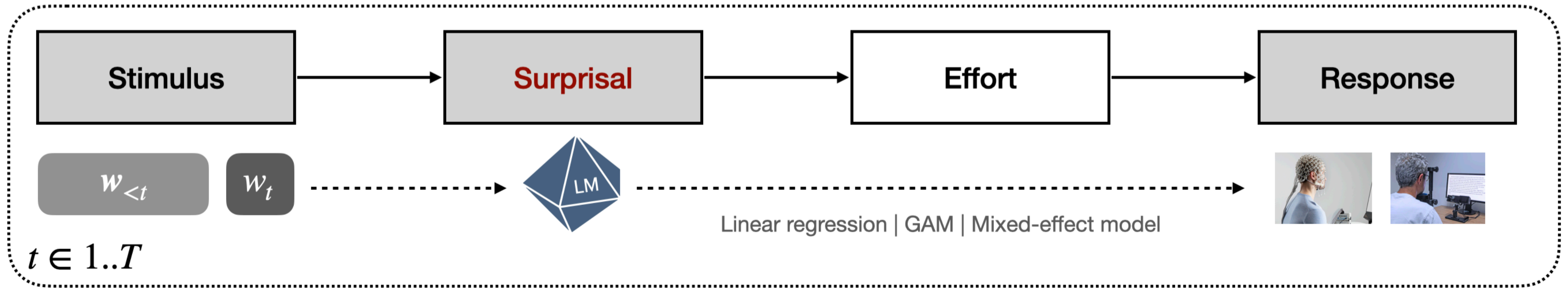


$$\iota^{\text{Surprisal}}(w_t; \boldsymbol{w}_{<t}) := -\log p(w_t \mid \boldsymbol{w}_{<t}) = -\log \sum_{v\in\Sigma} p(v \mid \boldsymbol{w}_{<t}) \boxed{\mathbb{1}\{v = w_t\}}$$

$$-\log \sum_{v\in\Sigma} p(v \mid \boldsymbol{w}_{<t}) \boxed{s_{\boldsymbol{w}_{<t}}\left(v, w_t\right)} =: \iota^{\text{Similarity-Adjusted Surprisal}}(w_t; \boldsymbol{w}_{<t})$$

Similarity-adjusted Surprisal





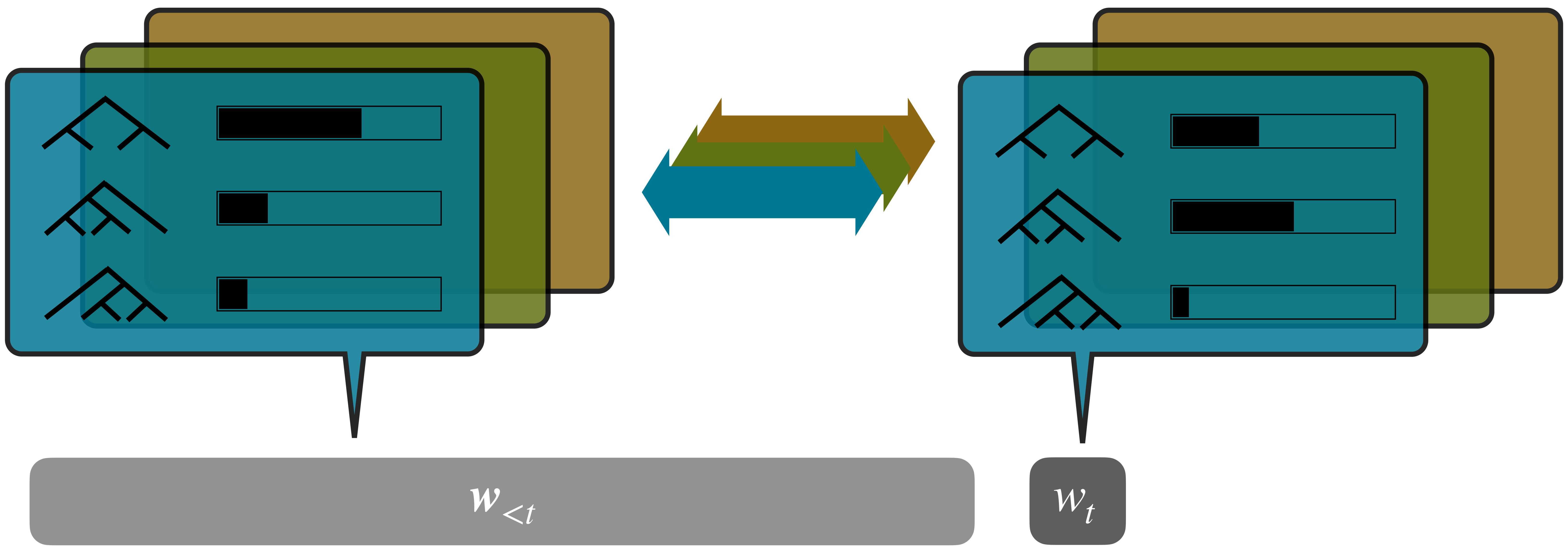
Language model: GPT-2 Small
Stimuli: English texts (Brown, Dundee, Natural Stories, Provo)
Responses: Eye-tracked and self-paced reading times (avg. across subjects)

Meister, Giulianelli, Pimentel. EMNLP 2024.

	Similarity-adjusted surprisal			
	Non-contextual	Contextual	POS	Orthographic
Brown	2.78***	-0.02	2.22	1.29*
Dundee	1.44***	0.01	0.69	0.64*
Natural Stories	3.18***	0.31**	1.04*	0.79*
Provo	1.34***	0.05	1.49	0.82

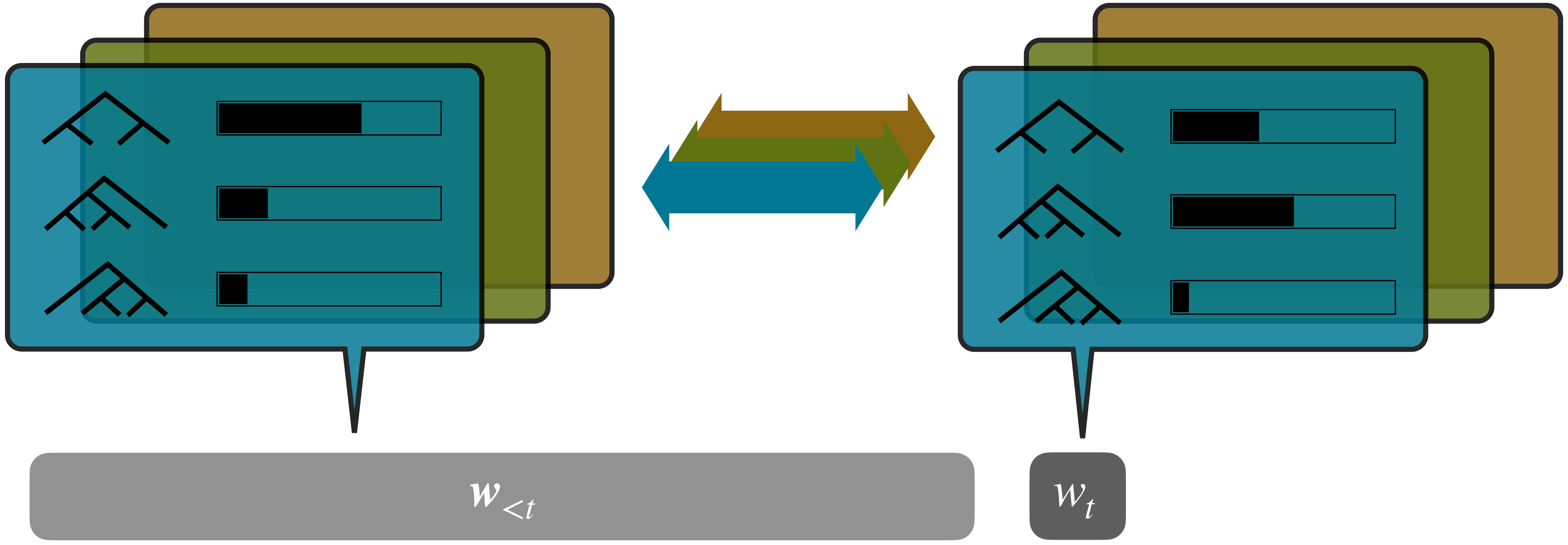
Δ_{LogLik} with respect to baseline model; predictors for current and previous three words.

Generalising Surprisal



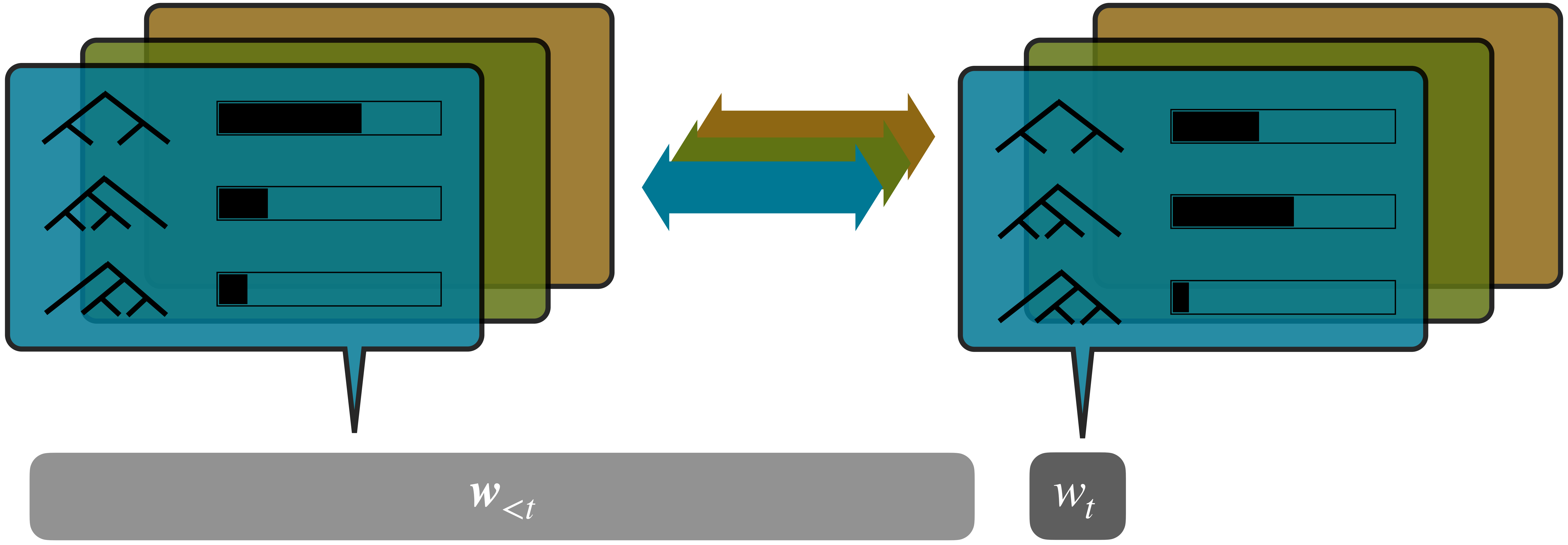
$$\iota^{\text{Surprisal}}(w_t; \boldsymbol{w}_{<t}) = -\log p(w_t \mid \boldsymbol{w}_{<t}) = -\log \sum_{v \in \Sigma} p(v \mid \boldsymbol{w}_{<t})\ \mathbf{1}\{v = w_t\}$$

Generalising Surprisal



$$\iota^{\text{Surprisal}}(w_t; \boldsymbol{w}_{<t}) = -\log p(w_t \mid \boldsymbol{w}_{<t}) = -\log \sum_{v \in \Sigma} p(v \mid \boldsymbol{w}_{<t})\, \mathbf{1}\{v = w_t\} = -\log \left(\mathbb{E}_{v \sim p(\cdot \mid \boldsymbol{w}_{<t})}\, \mathbf{1}\{v = w_t\} \right)$$

Generalising Surprisal



$$\iota^{\text{Surprisal}}(w_t; \boldsymbol{w}_{<t}) = -\log p(w_t \mid \boldsymbol{w}_{<t}) = -\log \sum_{v\in\Sigma} p(v \mid \boldsymbol{w}_{<t})\, \mathbf{1}\{v = w_t\} = \boxed{-\log}\left(\mathbb{E}_{v\sim p(\cdot\mid \boldsymbol{w}_{<t})} \boxed{\mathbf{1}\{v = w_t\}}\right)$$

$$\mathrm{f}\left(\mathbb{E}_{v\sim p(\cdot\mid \boldsymbol{w}_{<t})}\, \mathrm{g}\left(v, w_t, \boldsymbol{w}_{<t}\right)\right)$$

Generalised Surprisal

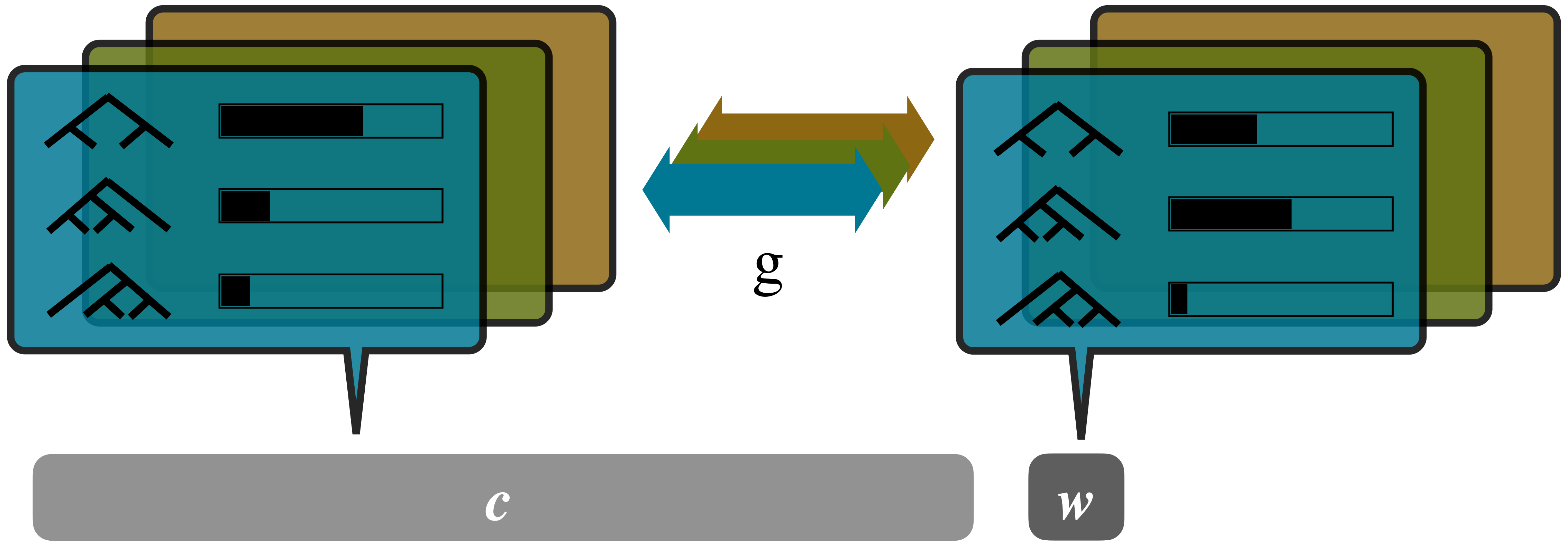
Giulianelli, Opedal, Cotterell. EMNLP 2024.

A generalised surprisal model is the pair (f, g) of

- a **warping function** $\mathrm{f} : \mathbb{R} \to \mathbb{R}$
- and a **scoring function** $\mathrm{g} : \Sigma^* \times \Sigma^* \times \Sigma^* \to \mathbb{R}$

Under a model (f, g), the **generalised surprisal** of a string $\boldsymbol{w}$ in a context $\boldsymbol{c}$ is

$$\iota_p^{(\mathrm{f},\mathrm{g})}(\boldsymbol{w}; \boldsymbol{c}) := \mathrm{f}\left(\mathbb{E}_{\boldsymbol{v} \sim p(\cdot|\boldsymbol{c})}\ \mathrm{g}(\boldsymbol{v}, \boldsymbol{w}, \boldsymbol{c})\right)$$



Generalised Surprisal

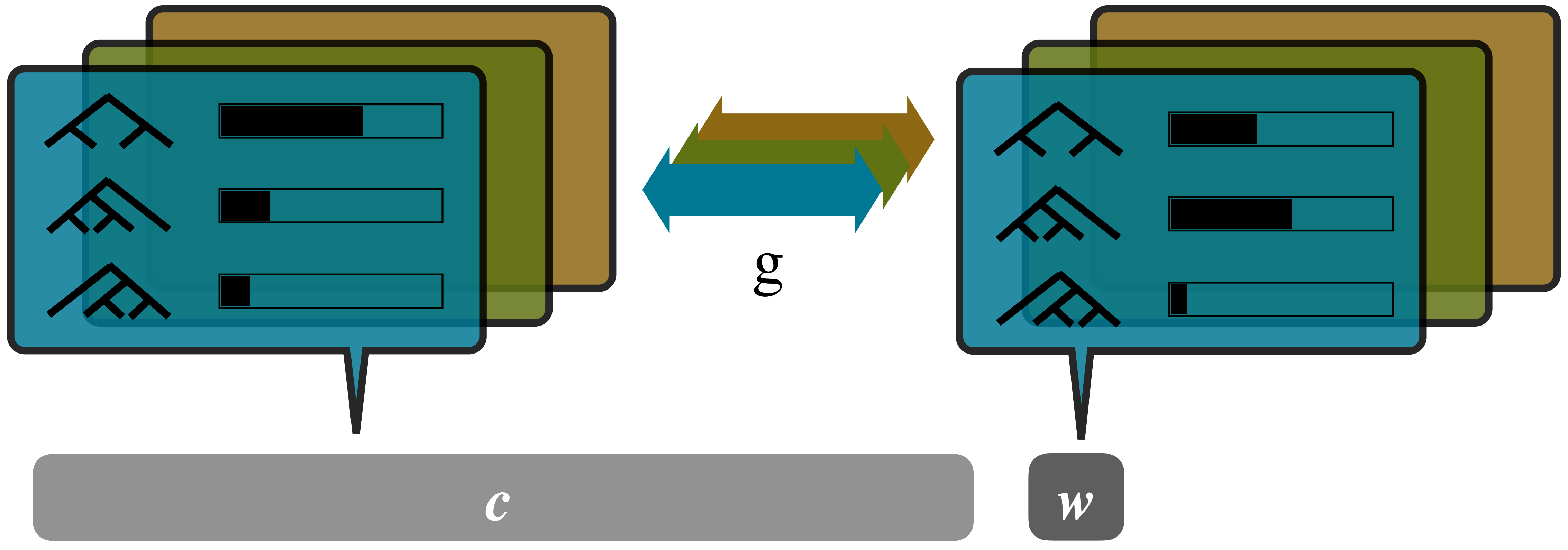
Giulianelli, Opedal, Cotterell. EMNLP 2024.

A generalised surprisal model is the pair (f, g) of

- a **warping function** $\mathrm{f} : \mathbb{R} \to \mathbb{R}$
- and a **scoring function** $\mathrm{g} : \Sigma^* \times \Sigma^* \times \Sigma^* \to \mathbb{R}$

Under a model (f, g), the **generalised surprisal** of a string $\boldsymbol{w}$ in a context $\boldsymbol{c}$ is

$$\iota_p^{(\mathrm{f},\mathrm{g})}(\boldsymbol{w};\boldsymbol{c}) := \mathrm{f}\left(\mathbb{E}_{\boldsymbol{v}\sim p(\cdot|\boldsymbol{c})}\ \mathrm{g}(\boldsymbol{v},\boldsymbol{w},\boldsymbol{c})\right)$$



Stimulus → $\iota_p^{(\mathrm{f},\mathrm{g})}(\boldsymbol{w};\boldsymbol{c})$ → Effort → Response

$\boldsymbol{w}_{<t}$ w_t ⇢

$+\ \mathrm{g}$ ⇢ f

$t \in 1..T$

Generalised Surprisal
Warping Functions: Surprisal v. Probability

A generalised surprisal model is the pair (f, g) of

- a **warping function** $\mathrm{f} : \mathbb{R} \to \mathbb{R}$
- and a **scoring function** $\mathrm{g} : \Sigma^* \times \Sigma^* \times \Sigma^* \to \mathbb{R}$

Under a model (f, g), the **generalised surprisal** of a string $\boldsymbol{w}$ in a context $\boldsymbol{c}$ is

$$\iota_p^{(\mathrm{f},\mathrm{g})}(\boldsymbol{w}; \boldsymbol{c}) := \mathrm{f}\left(\mathbb{E}_{\boldsymbol{v}\sim p(\cdot|\boldsymbol{c})}\ \mathrm{g}(\boldsymbol{v}, \boldsymbol{w}, \boldsymbol{c})\right)$$

Next-word Surprisal	Next-word Probability
$\mathrm{f}(x) = -\log(x)$	$\mathrm{f}(x) = x$
$\mathrm{g}(v, w_t, \boldsymbol{w}_{<t}) = \mathbf{1}\{v = w_t\}$	$\mathrm{g}(v, w_t, \boldsymbol{w}_{<t}) = \mathbf{1}\{v = w_t\}$

Generalised Surprisal
Warping Functions: Surprisal v. Probability

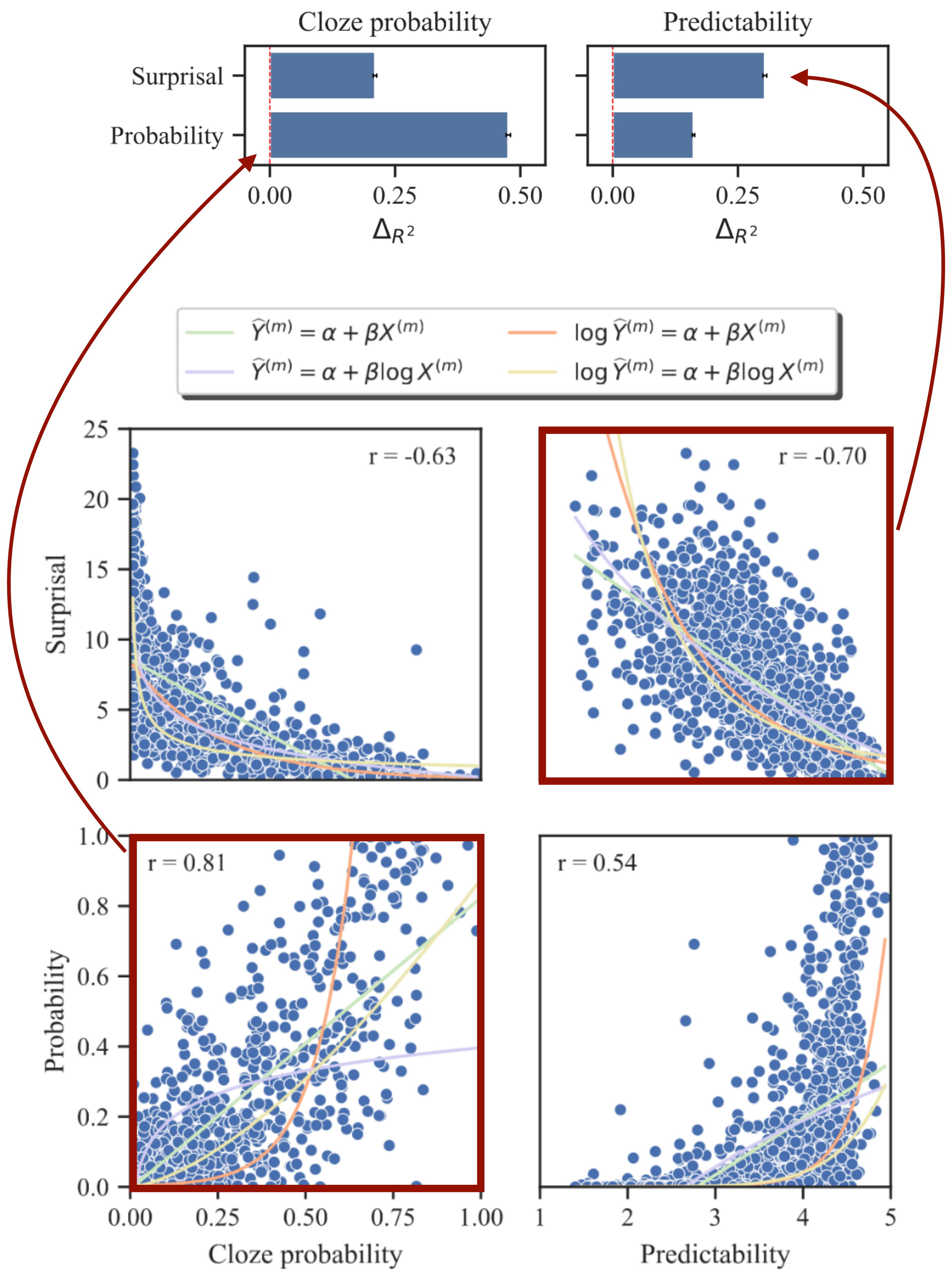
A generalised surprisal model is the pair (f, g) of

- a **warping function** $\mathrm{f} : \mathbb{R} \rightarrow \mathbb{R}$
- and a **scoring function** $\mathrm{g} : \Sigma^* \times \Sigma^* \times \Sigma^* \rightarrow \mathbb{R}$

Under a model (f, g), the **generalised surprisal** of a string $\boldsymbol{w}$ in a context $\boldsymbol{c}$ is

$$\iota_p^{(\mathrm{f},\mathrm{g})}(\boldsymbol{w};\boldsymbol{c}) := \mathrm{f}\left(\mathbb{E}_{v\sim p(\cdot|\boldsymbol{c})}\ \mathrm{g}(v, \boldsymbol{w}, \boldsymbol{c})\right)$$

Next-word Surprisal	Next-word Probability
$\mathrm{f}(x) = -\log(x)$	$\mathrm{f}(x) = x$
$\mathrm{g}(v, w_t, \boldsymbol{w}_{<t}) = \mathbf{1}\{v = w_t\}$	$\mathrm{g}(v, w_t, \boldsymbol{w}_{<t}) = \mathbf{1}\{v = w_t\}$



Language model: GPT-2 Small
Stimuli: M = 1726 target–context pairs from English novels (de Varda et al. 2023)
Responses: Cloze probability (smoothed); predictability ratings (1 to 5)

Giulianelli, Opedal, Cotterell. EMNLP 2024.

Generalised Surprisal

Scoring Functions: Information Value

A generalised surprisal model is the pair (f, g) of

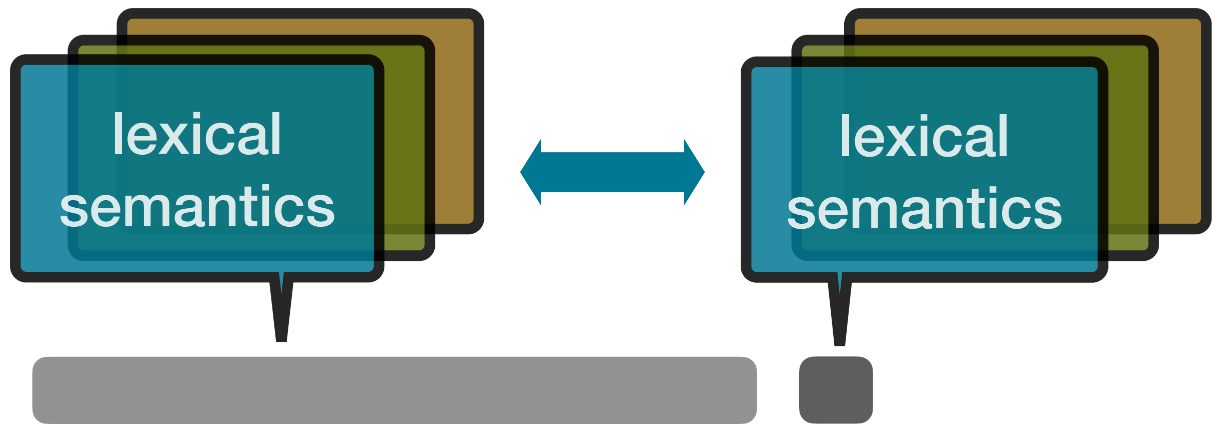
- a **warping function** $\mathrm{f} : \mathbb{R} \to \mathbb{R}$
- and a **scoring function** $\mathrm{g} : \Sigma^* \times \Sigma^* \times \Sigma^* \to \mathbb{R}$

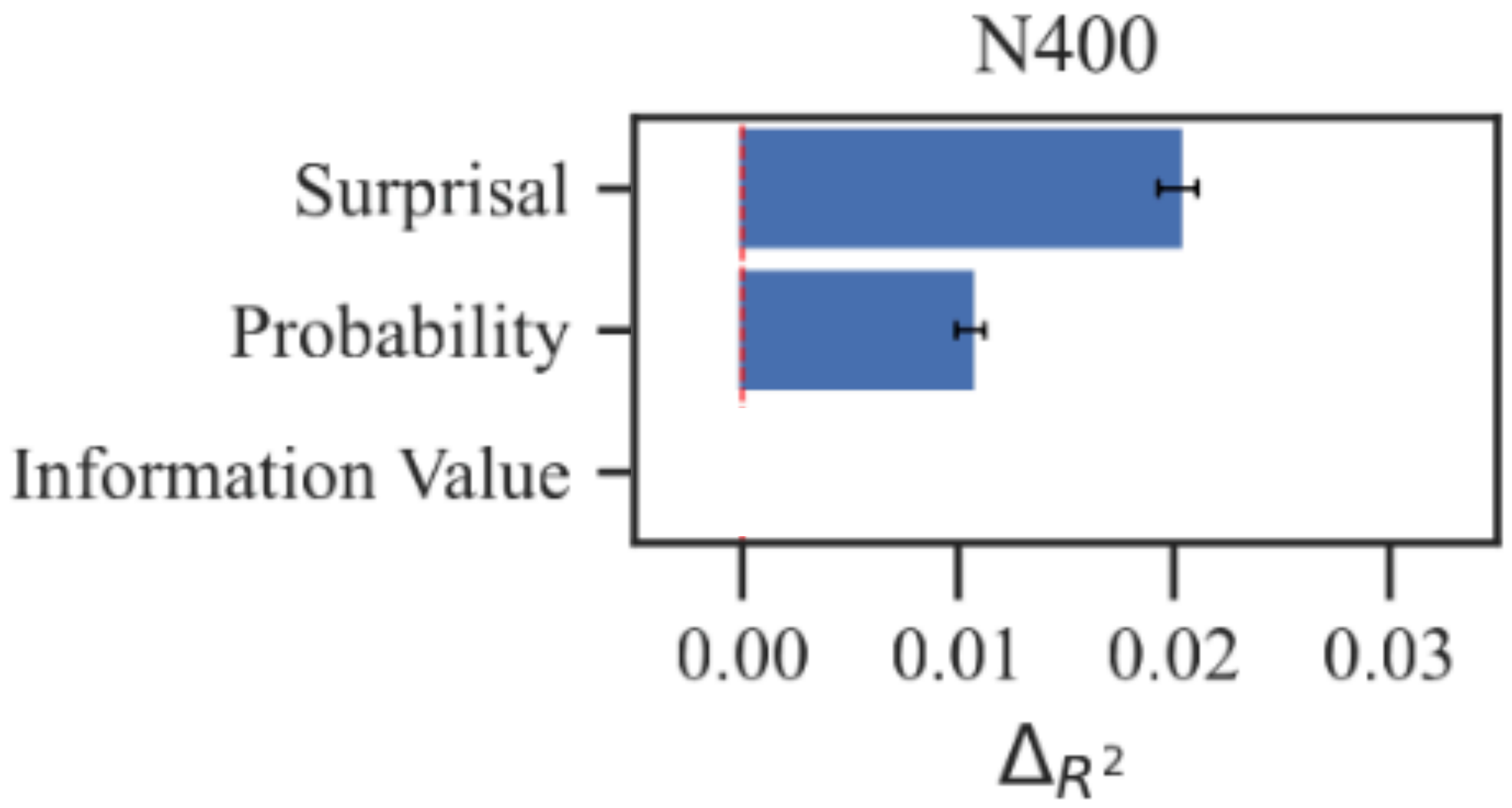
Under a model (f, g), the **generalised surprisal** of a string $\boldsymbol{w}$ in a context $\boldsymbol{c}$ is

$$\iota_p^{(\mathrm{f},\mathrm{g})}(\boldsymbol{w};\boldsymbol{c}) := \mathrm{f}\left(\mathbb{E}_{\boldsymbol{v}\sim p(\cdot|\boldsymbol{c})}\ \mathrm{g}(\boldsymbol{v},\boldsymbol{w},\boldsymbol{c})\right)$$

Next-word Information Value
$\mathrm{f}(x) = x$
$\mathrm{g}(v, w_t, \boldsymbol{w}_{<t}) = d_{\boldsymbol{w}_{<t}}(v, w_t)$

Giulianelli, Wallbridge, Fernández. EMNLP 2023.





Language model: GPT-2 Small
$d_{\boldsymbol{w}_{<t}}(v, w_t) \to$ cosine between contextualised word embeddings
Stimuli: M = 1726 target–context pairs from English novels (de Varda et al. 2023)
Response: N400 (avg. across participants)

Giulianelli, Opedal, Cotterell. EMNLP 2024.

Generalised Surprisal
Scoring Functions: Information Value

A generalised surprisal model is the pair (f, g) of

- a **warping function** $\mathrm{f} : \mathbb{R} \to \mathbb{R}$
- and a **scoring function** $\mathrm{g} : \Sigma^* \times \Sigma^* \times \Sigma^* \to \mathbb{R}$

Under a model (f, g), the **generalised surprisal** of a string $\boldsymbol{w}$ in a context $\boldsymbol{c}$ is

$$\iota_p^{(\mathrm{f},\mathrm{g})}(\boldsymbol{w};\boldsymbol{c}) := \mathrm{f}\left(\mathbb{E}_{\boldsymbol{v}\sim p(\cdot|\boldsymbol{c})}\ \mathrm{g}(\boldsymbol{v},\boldsymbol{w},\boldsymbol{c})\right)$$

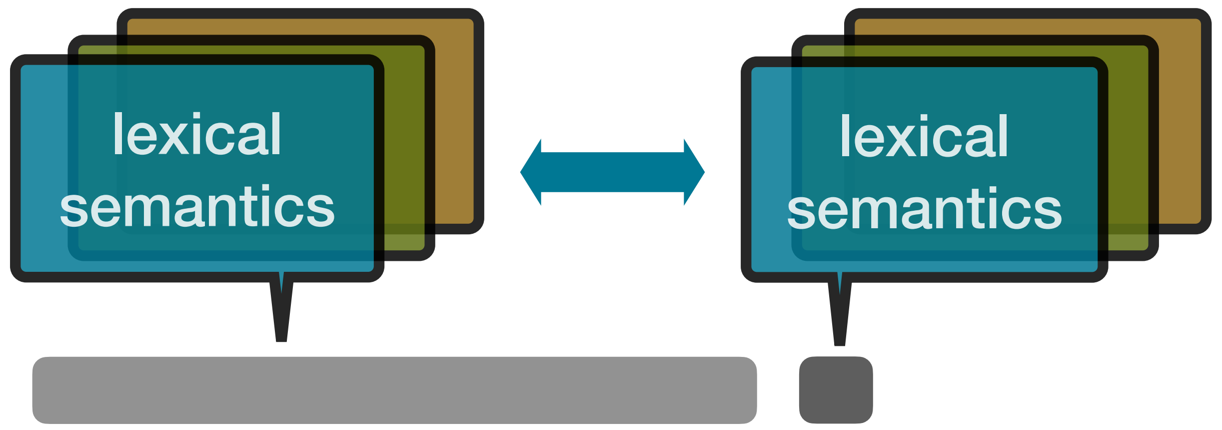
Next-word Information Value

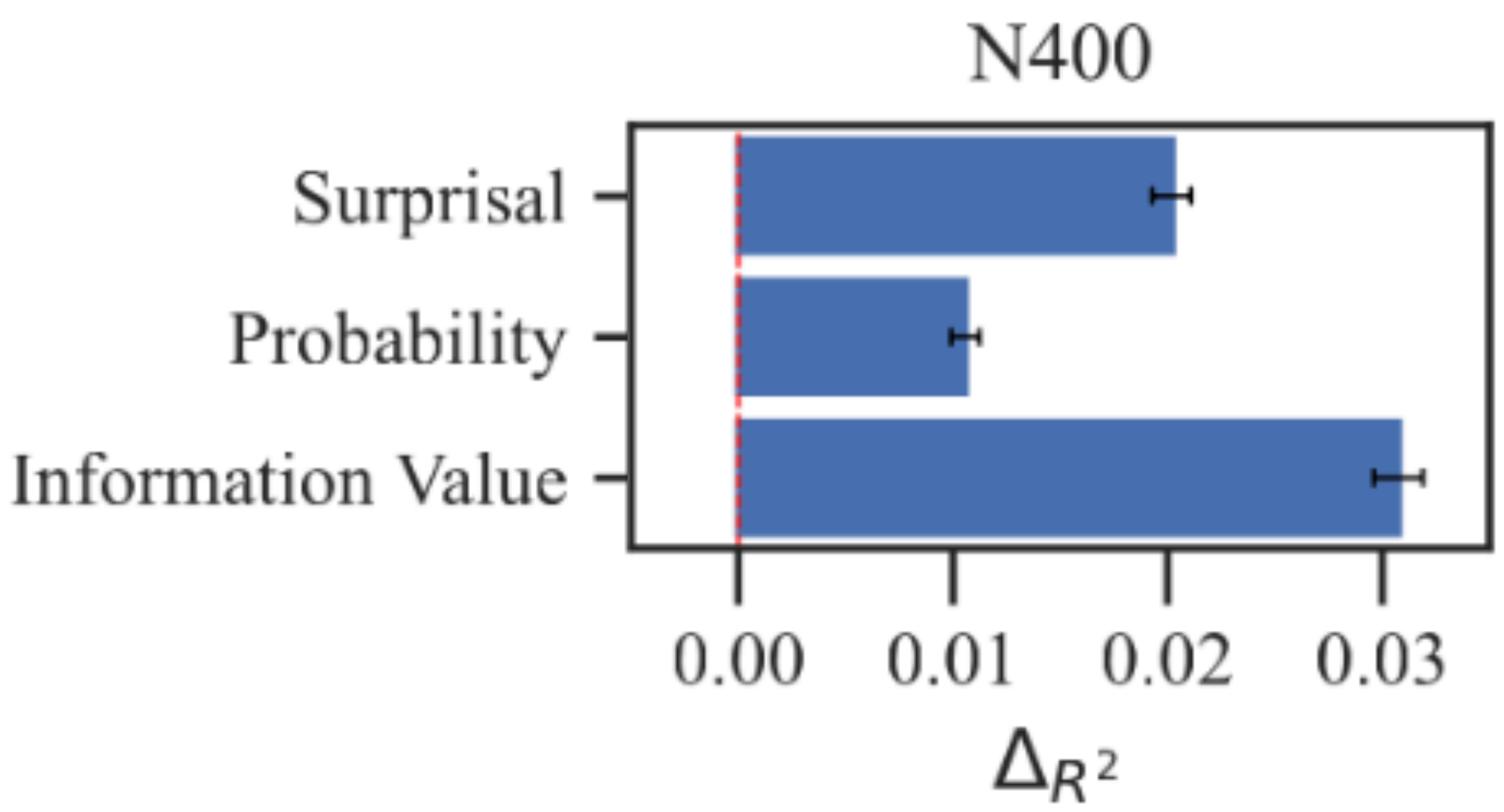
$$\mathrm{f}(x) = x$$

$$\mathrm{g}(v, w_t, \boldsymbol{w}_{<t}) = d_{\boldsymbol{w}_{<t}}(v, w_t)$$

$d_{w_{<t}}(v, w_t) \to$ cosine between contextualised word embeddings

Giulianelli, Wallbridge, Fernández. EMNLP 2023.





Language model: GPT-2 Small
$d_{w_{<t}}(v, w_t) \to$ cosine between contextualised word embeddings
Stimuli: M = 1726 target–context pairs from English novels (de Varda et al. 2023)
Response: N400 (avg. across participants)

Giulianelli, Opedal, Cotterell. EMNLP 2024.

Generalised Surprisal

Responsive Uncertainty

A generalised surprisal model is the pair (f, g) of

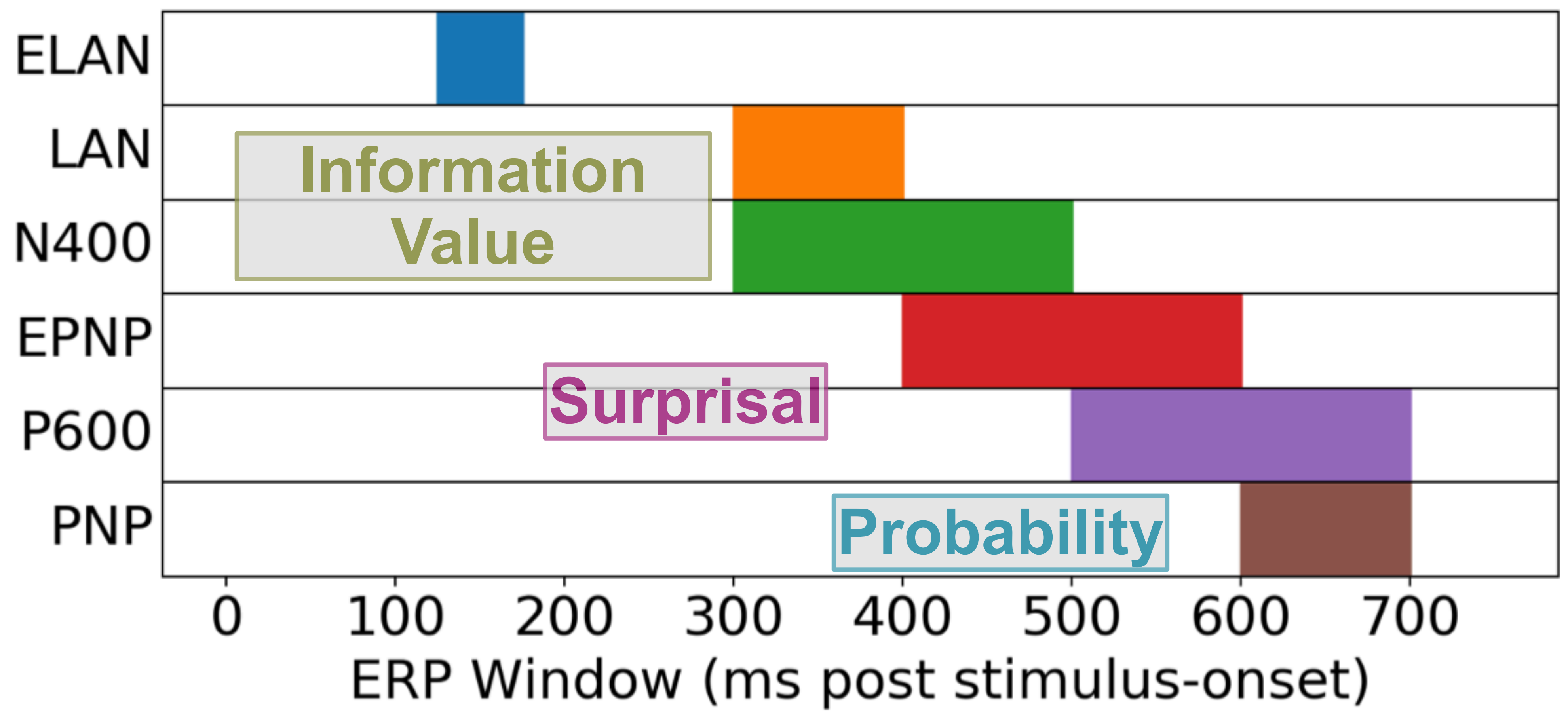
- ✦ a **warping function** $\mathrm{f} : \mathbb{R} \to \mathbb{R}$
- ✦ and a **scoring function** $\mathrm{g} : \Sigma^* \times \Sigma^* \times \Sigma^* \to \mathbb{R}$

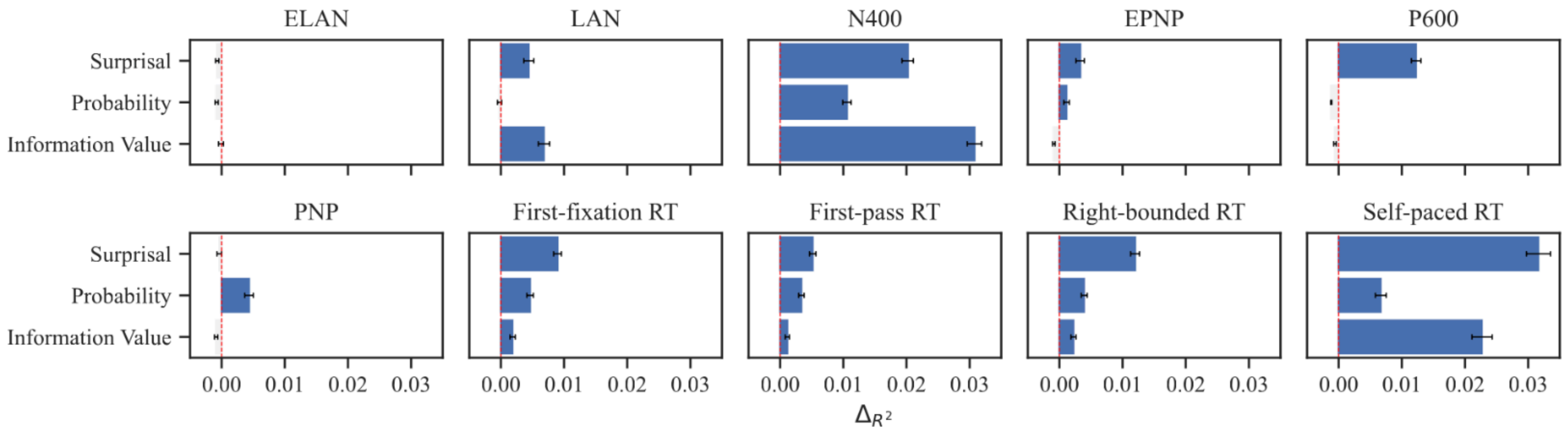
Under a model (f, g), the **generalised surprisal** of a string $\boldsymbol{w}$ in a context $\boldsymbol{c}$ is

$$\iota_p^{(\mathrm{f},\mathrm{g})}(\boldsymbol{w};\boldsymbol{c}) := \mathrm{f}\left(\mathbb{E}_{\boldsymbol{v}\sim p(\cdot|\boldsymbol{c})}\ \mathrm{g}(\boldsymbol{v},\boldsymbol{w},\boldsymbol{c})\right)$$

Surprisal	Probability	Information Value
$\mathrm{f}(x) = -\log(x)$	$\mathrm{f}(x) = x$	$\mathrm{f}(x) = x$
$\mathrm{g}(v, w_t, \boldsymbol{w}_{<t}) = \mathbf{1}\{v = w_t\}$	$\mathrm{g}(v, w_t, \boldsymbol{w}_{<t}) = \mathbf{1}\{v = w_t\}$	$\mathrm{g}(v, w_t, \boldsymbol{w}_{<t}) = d_{\boldsymbol{w}_{<t}}(v, w_t)$

$d_{\boldsymbol{w}_{<t}}(v, w_t) \to$ cosine between contextualised word embeddings





Language model: GPT-2 Small
Stimuli: M = 1726 target–context pairs from English novels (de Varda et al. 2023)
Response: ELAN, LAN, N400, EPNP, P600, PNP (avg. across participants)

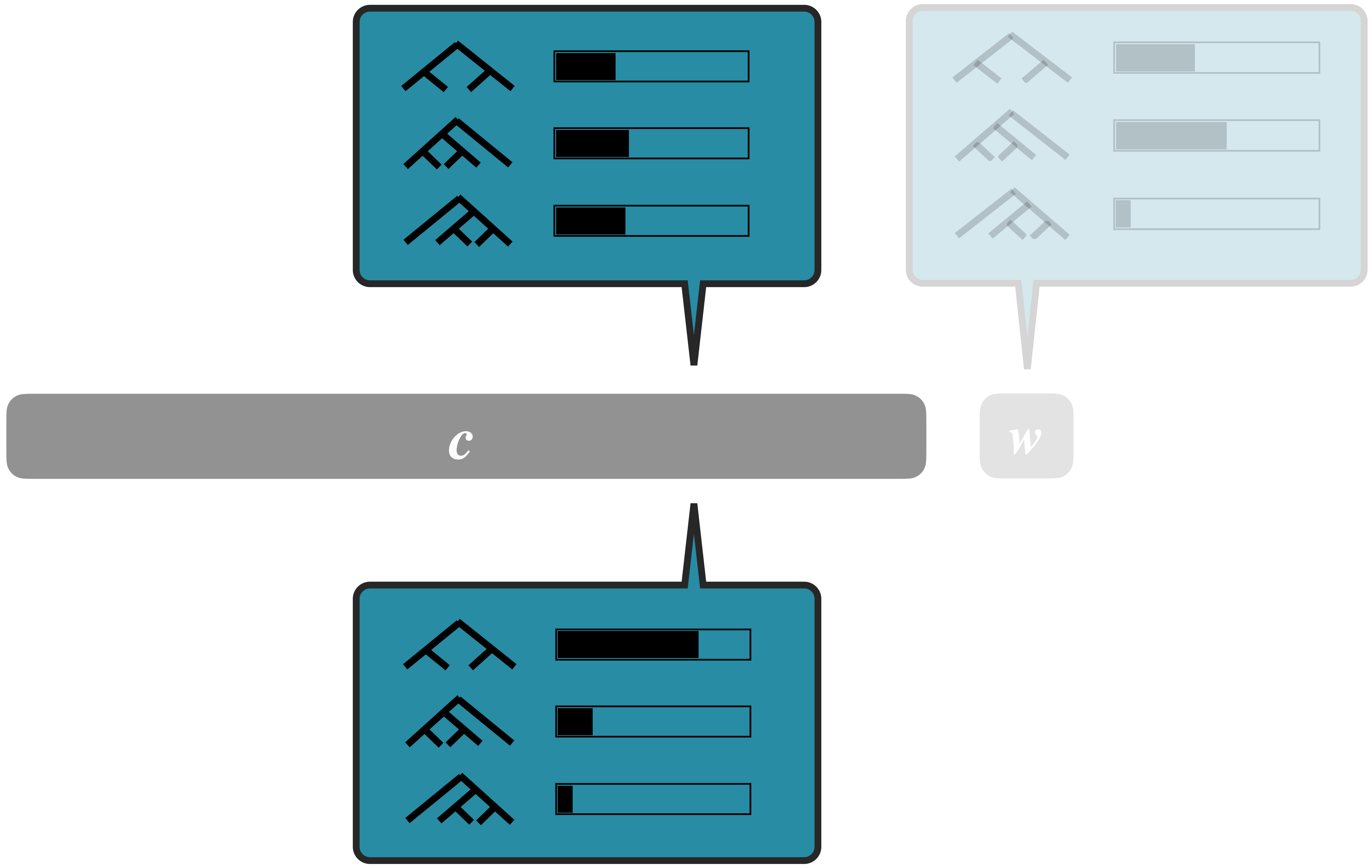
Giulianelli, Opedal, Cotterell. EMNLP 2024.

Generalised Surprisal
Anticipatory Uncertainty

$$\iota_p^{(\mathrm{f},\mathrm{g})}(\boldsymbol{w};\boldsymbol{c}) := \mathrm{f}\left(\mathbb{E}_{\boldsymbol{v}\sim p(\cdot|\boldsymbol{c})}\ \mathrm{g}(\boldsymbol{v},\boldsymbol{w},\boldsymbol{c})\right)$$

We call a generalised surprisal model (f,g) **anticipatory** if the scoring function g is constant in $\boldsymbol{w}$, i.e., if $\forall \boldsymbol{v},\boldsymbol{w},\boldsymbol{w}',\boldsymbol{c} \in \Sigma^* : \mathrm{g}(\boldsymbol{v},\boldsymbol{w},\boldsymbol{c}) = \mathrm{g}(\boldsymbol{v},\boldsymbol{w}',\boldsymbol{c})$.

Otherwise, we call (f,g) **responsive**.



Generalised Surprisal
Anticipatory Uncertainty

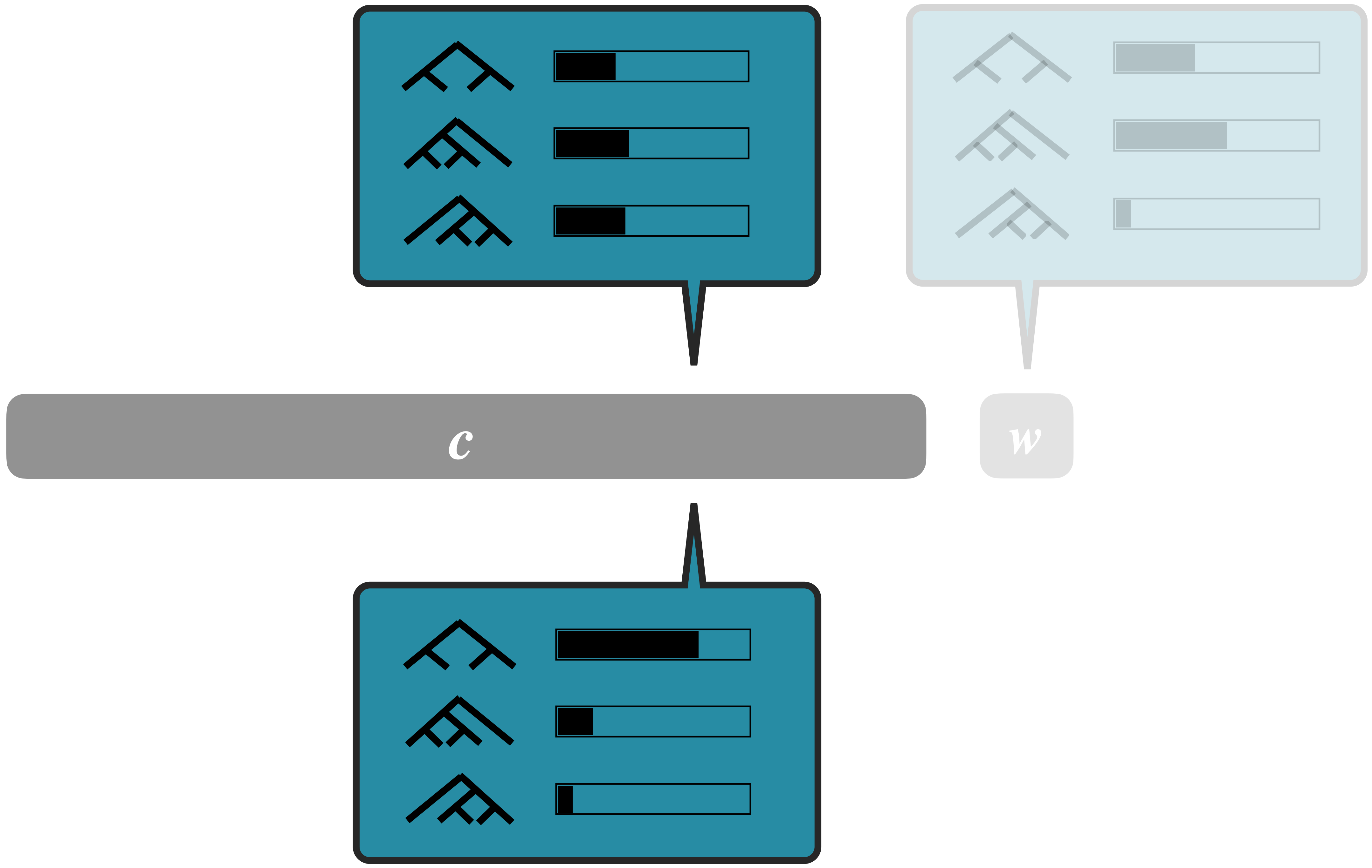
$$\iota_p^{(\mathrm{f},\mathrm{g})}(\boldsymbol{w};\boldsymbol{c}) := \mathrm{f}\left(\mathbb{E}_{\boldsymbol{v}\sim p(\cdot\mid\boldsymbol{c})}\ \mathrm{g}(\boldsymbol{v},\boldsymbol{w},\boldsymbol{c})\right)$$

We call a generalised surprisal model (f,g) **anticipatory** if the scoring function g is constant in $\boldsymbol{w}$, i.e., if $\forall \boldsymbol{v},\boldsymbol{w},\boldsymbol{w}',\boldsymbol{c}\in\Sigma^* : \mathrm{g}(\boldsymbol{v},\boldsymbol{w},\boldsymbol{c}) = \mathrm{g}(\boldsymbol{v},\boldsymbol{w}',\boldsymbol{c})$.

Otherwise, we call (f,g) **responsive**.

Next-word Entropy	
$\mathrm{f}(x) = x$	$\mathrm{g}(\boldsymbol{v},\boldsymbol{w},\boldsymbol{c}) = -\sum_{u\in\Sigma} \mathbf{1}\{u \preceq \boldsymbol{v}\}\log p(u \mid \boldsymbol{c})$

Sequence Entropy	
$\mathrm{f}(x) = x$	$\mathrm{g}(\boldsymbol{v},\boldsymbol{w},\boldsymbol{c}) = -\log p(\boldsymbol{v} \mid \boldsymbol{c})$



Generalised Surprisal
Anticipatory Uncertainty

$$\iota_p^{(\mathrm{f},\mathrm{g})}(\boldsymbol{w};\boldsymbol{c}) := \mathrm{f}\left(\mathbb{E}_{\boldsymbol{v}\sim p(\cdot\mid\boldsymbol{c})}\,\mathrm{g}(\boldsymbol{v},\boldsymbol{w},\boldsymbol{c})\right)$$

We call a generalised surprisal model (f,g) **anticipatory** if the scoring function g is constant in $\boldsymbol{w}$, i.e., if $\forall \boldsymbol{v},\boldsymbol{w},\boldsymbol{w}',\boldsymbol{c} \in \Sigma^* : \mathrm{g}(\boldsymbol{v},\boldsymbol{w},\boldsymbol{c}) = \mathrm{g}(\boldsymbol{v},\boldsymbol{w}',\boldsymbol{c})$.

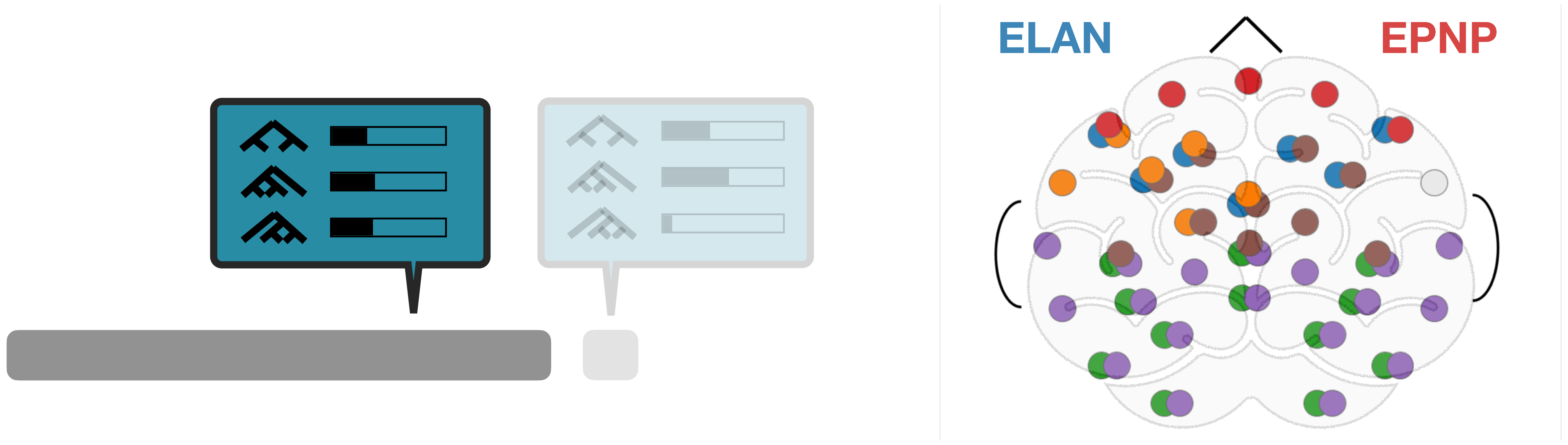
Otherwise, we call (f,g) **responsive**.

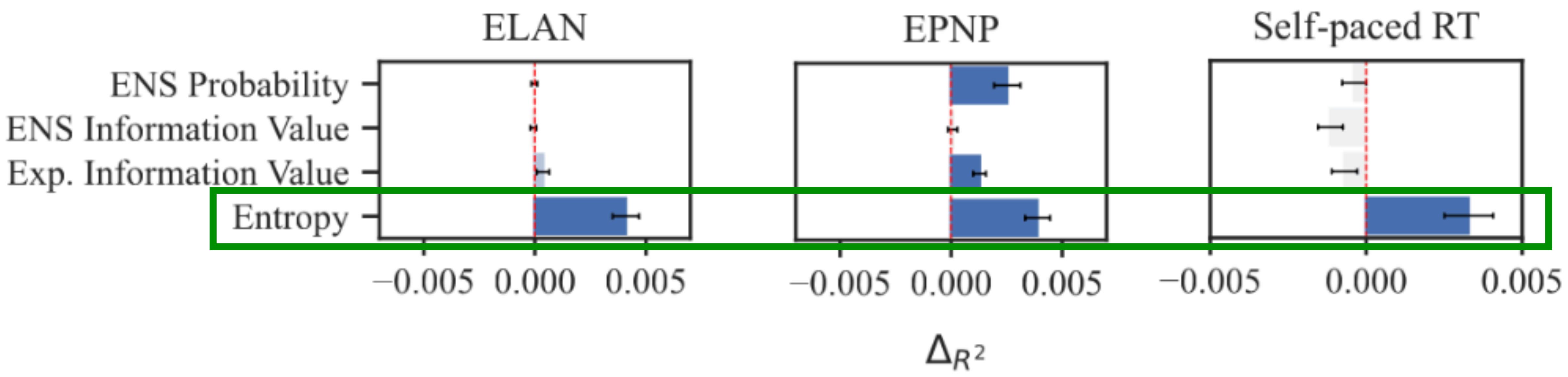
Next-word Entropy

$$\mathrm{f}(x) = x \qquad \mathrm{g}(\boldsymbol{v},\boldsymbol{w},\boldsymbol{c}) = -\sum_{u\in\Sigma} \mathbf{1}\{u \preceq \boldsymbol{v}\}\log p(u \mid \boldsymbol{c})$$

Sequence Entropy

$$\mathrm{f}(x) = x \qquad \mathrm{g}(\boldsymbol{v},\boldsymbol{w},\boldsymbol{c}) = -\log p(\boldsymbol{v} \mid \boldsymbol{c})$$





Increase in predictive power, Δ_{R^2}, over next-word entropy baseline.

Language model: GPT-2 Small
Stimuli: M = 1726 target–context pairs from English novels (de Varda et al. 2023)
Response: ELAN, EPNP; self-paced reading times

Giulianelli, Opedal, Cotterell. EMNLP 2024.

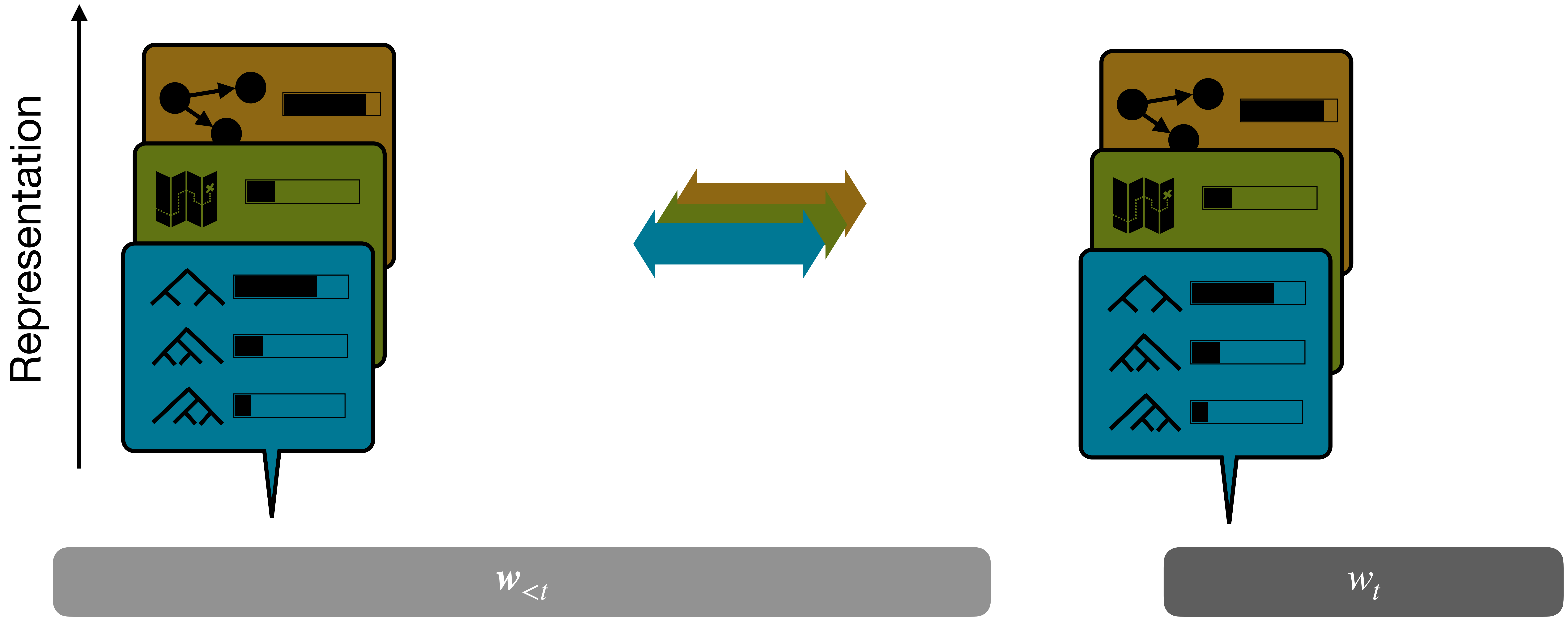
Generalised Surprisal

Summary: Most predictive uncertainty measure by response type

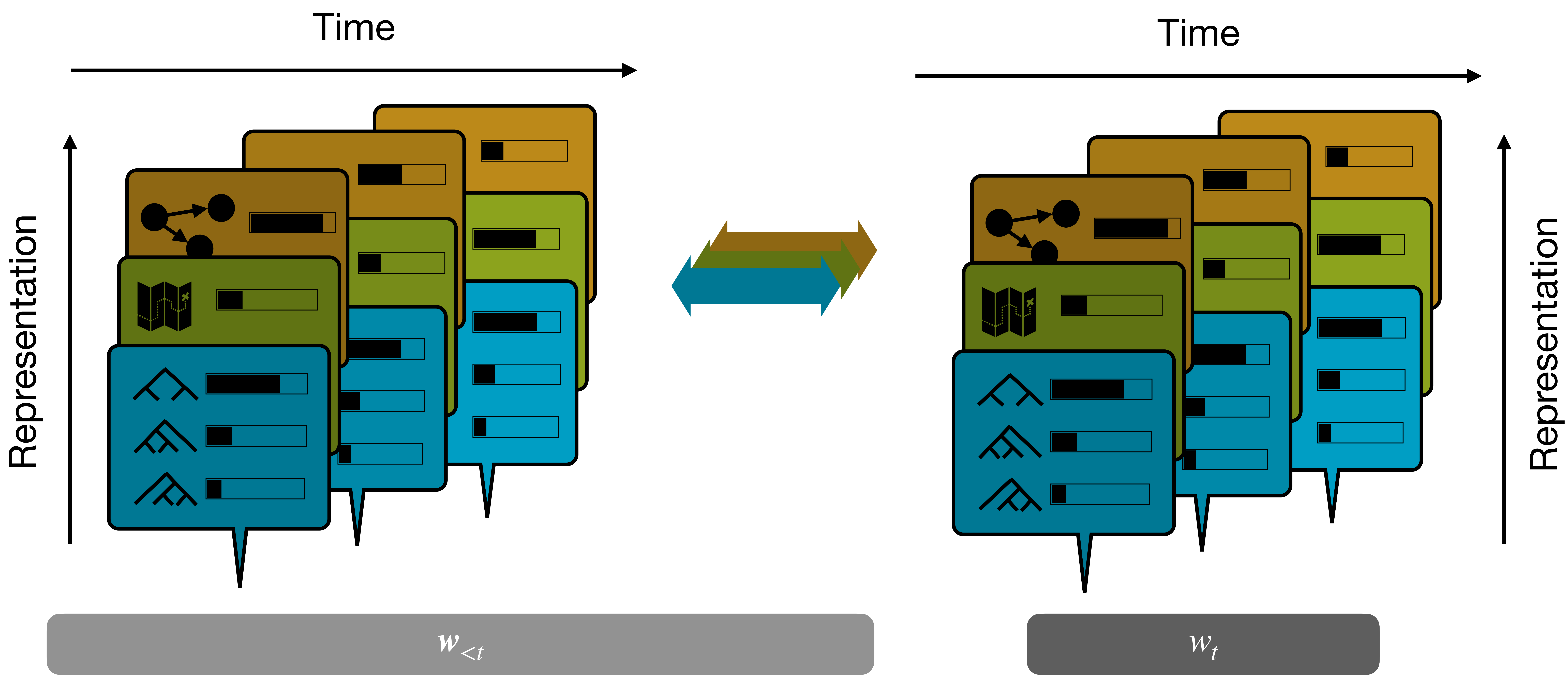
	ELAN	LAN	N400	EPNP	P600	PNP	First-fixation RT	First-pass RT	Right-bounded RT	Self-paced RT
Responsive		Information value	Information value	Surprisal	Surprisal	Probability	Surprisal	Surprisal	Surprisal	Surprisal
Anticipatory	(Sequence) Entropy	Information value	Information value	(Sequence) Entropy	(Sequence) Entropy	Exp. Next-symbol Probability	Exp. Next-symbol Information Value	Exp. Next-symbol Information Value	Exp. Next-symbol Information Value	(Sequence) Entropy

$\mathcal{M} = ($ p_{LM}, sampling procedure, **warping function**, **scoring function, anticipatory/responsive**)

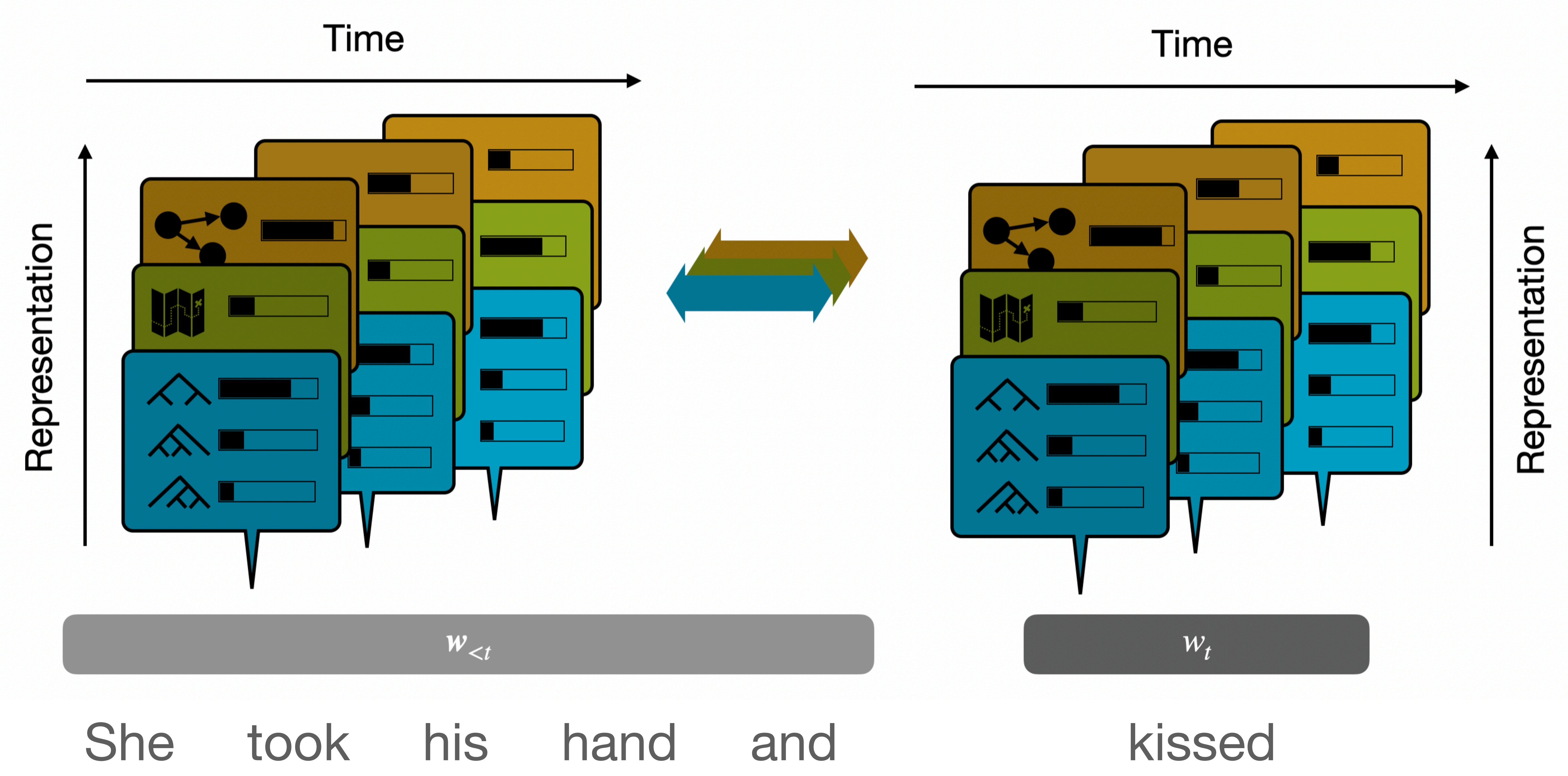
Incremental alternative sampling



Incremental alternative sampling

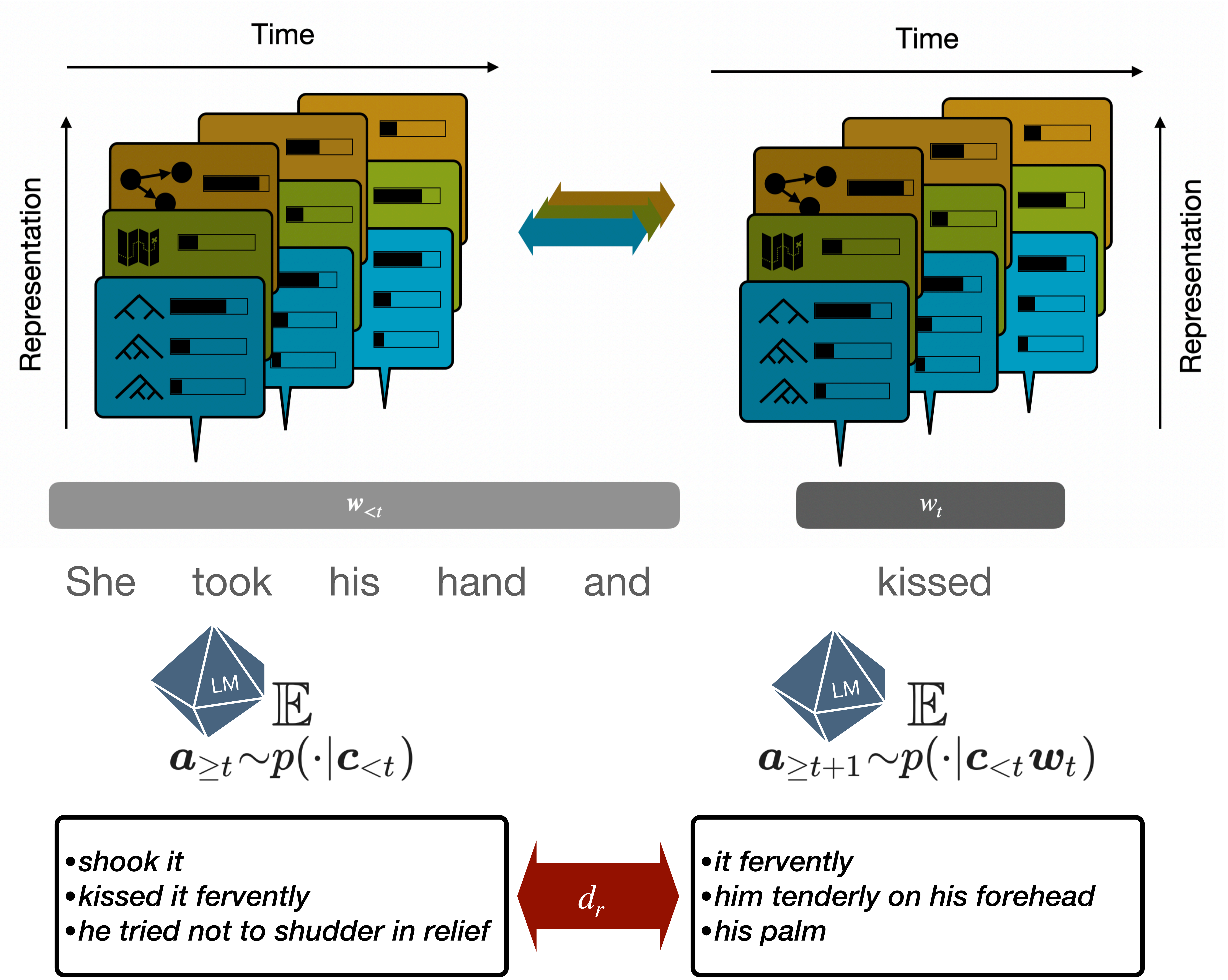


Incremental alternative sampling



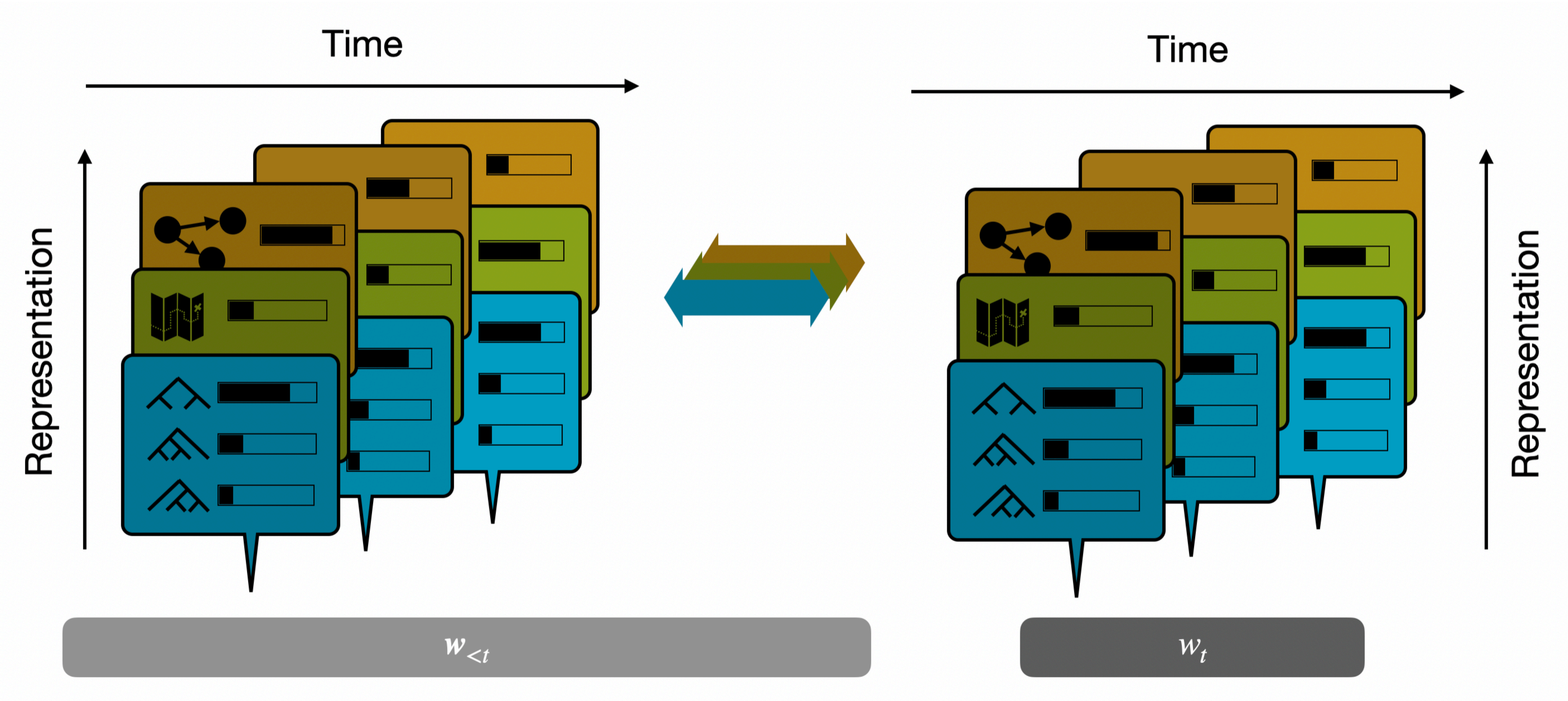
Giulianelli, Wallbridge, Cotterell, Fernández. *Journal of Memory and Language.* 2026.

Incremental alternative sampling



Giulianelli, Wallbridge, Cotterell, Fernández. *Journal of Memory and Language.* 2026.

Incremental alternative sampling



She took his hand and kissed

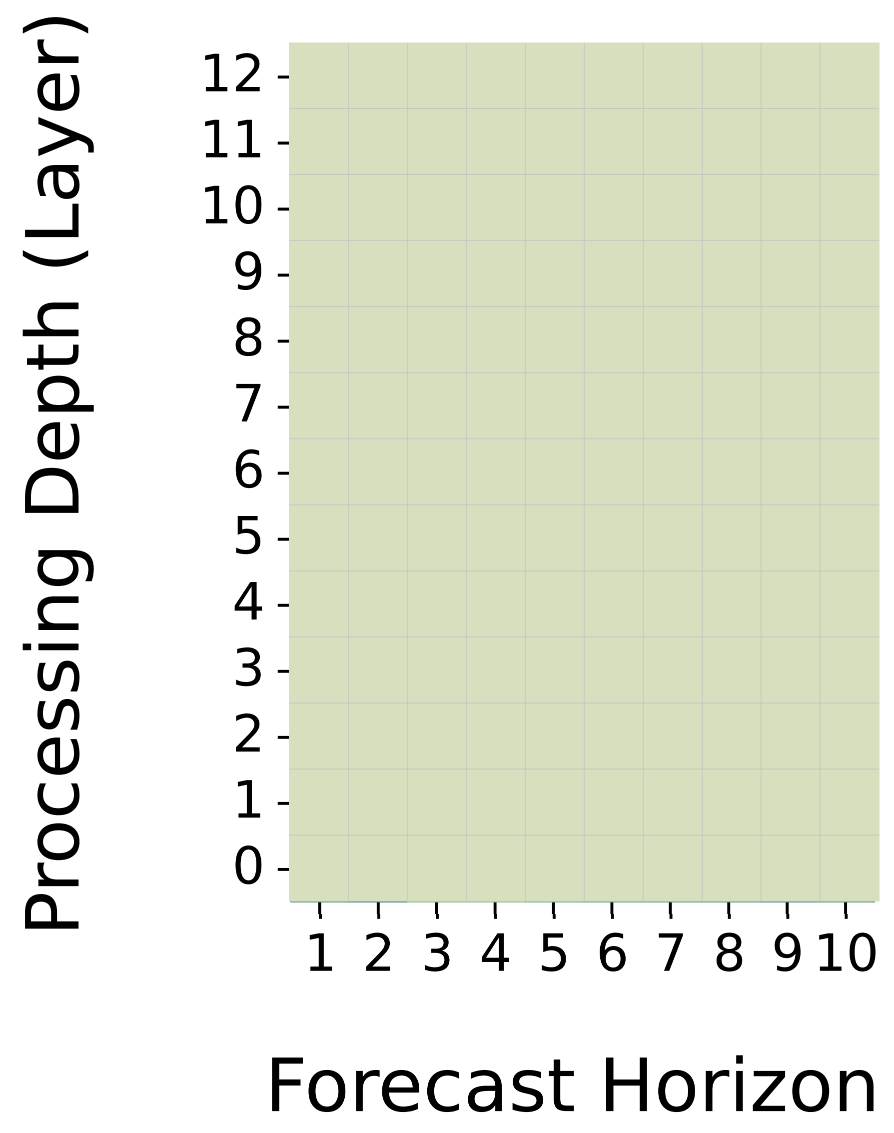
$\mathbb{E}_{\boldsymbol{a}_{\geq t} \sim p(\cdot | \boldsymbol{c}_{<t})}$

$\mathbb{E}_{\boldsymbol{a}_{\geq t+1} \sim p(\cdot | \boldsymbol{c}_{<t} \boldsymbol{w}_t)}$

- *shook it*
- *kissed it fervently*
- *he tried not to shudder in relief*

d_r

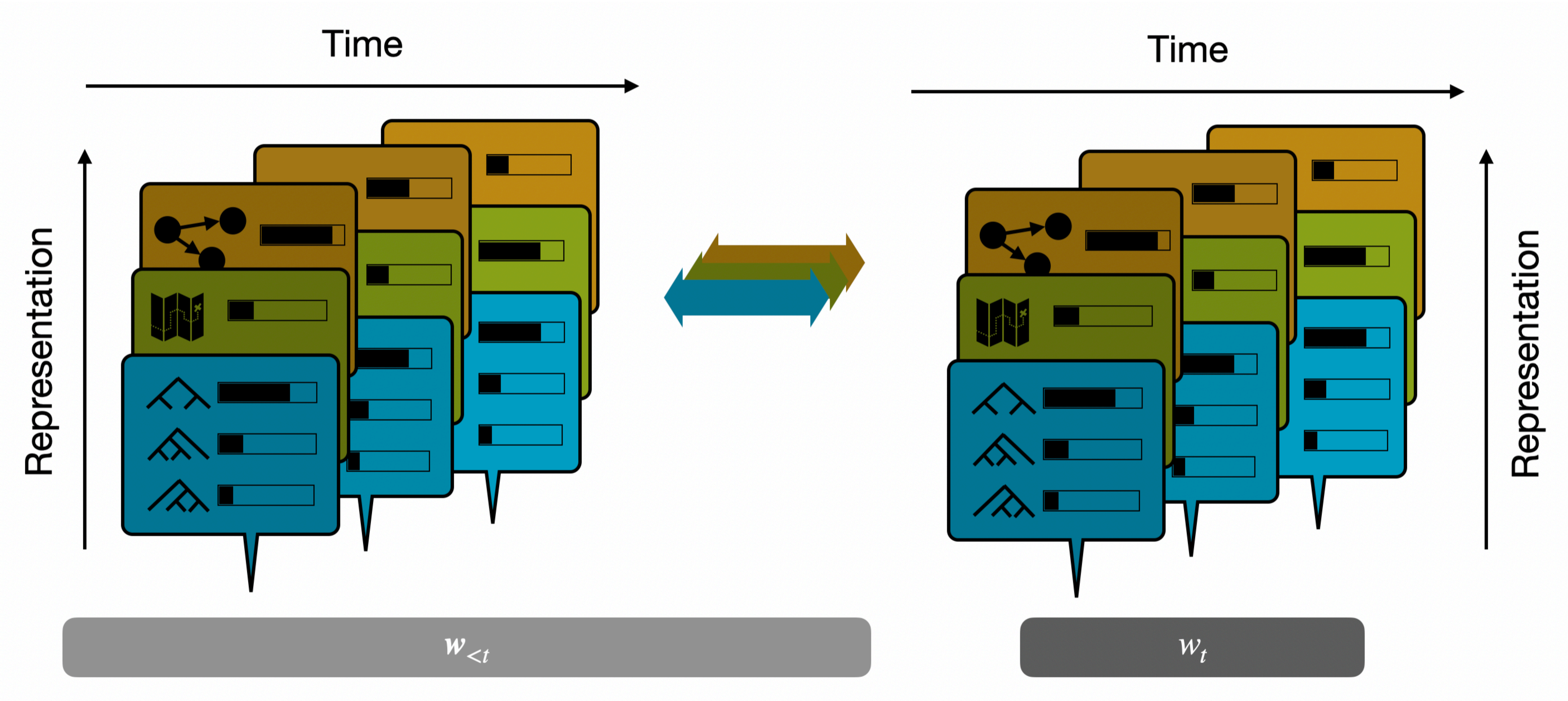
- *it fervently*
- *him tenderly on his forehead*
- *his palm*



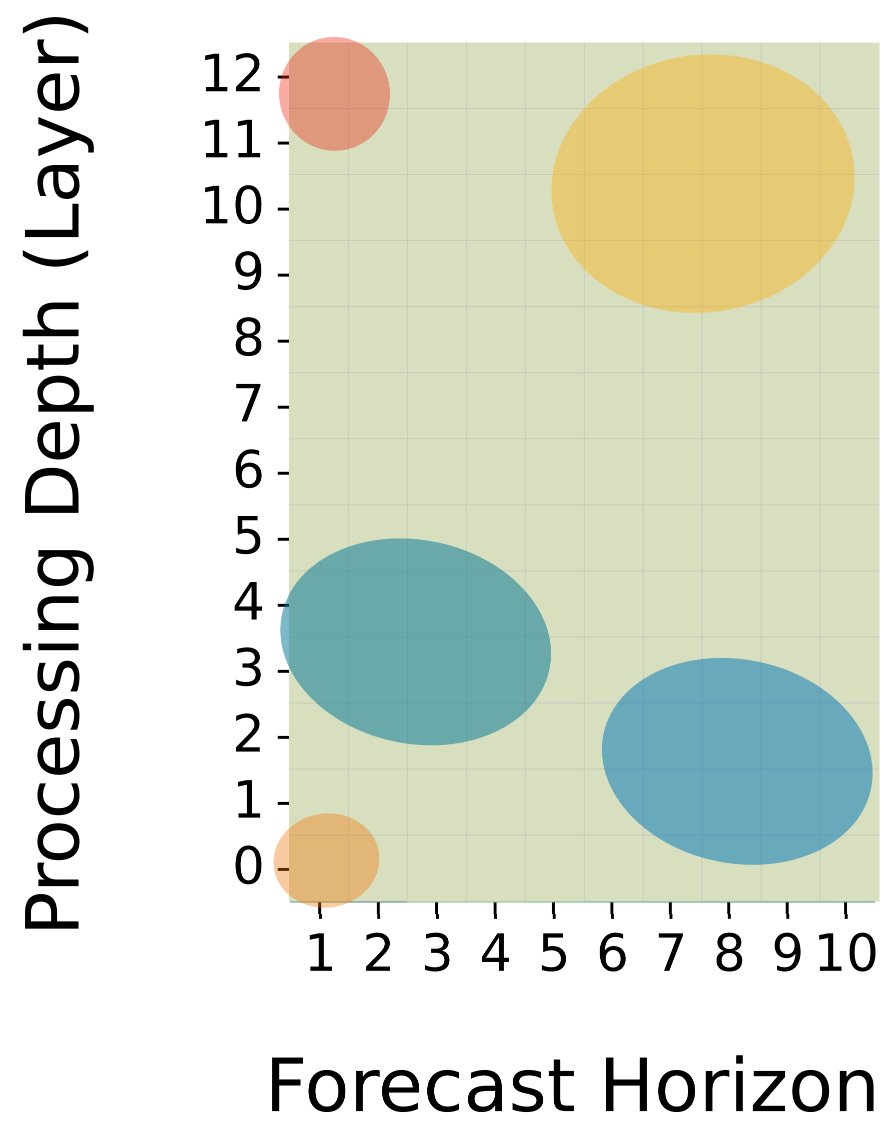
$$\mathrm{d}_r \left(\boldsymbol{a}_{\geq t} , \boldsymbol{w}_t \boldsymbol{a}_{\geq t+1} \right)$$

Giulianelli, Wallbridge, Cotterell, Fernández. *Journal of Memory and Language.* 2026.

Incremental alternative sampling



She took his hand and kissed



$$\mathbb{E}_{\boldsymbol{a}_{\geq t} \sim p(\cdot|\boldsymbol{c}_{<t})}$$

$$\mathbb{E}_{\boldsymbol{a}_{\geq t+1} \sim p(\cdot|\boldsymbol{c}_{<t}\boldsymbol{w}_t)}$$

$$\mathrm{d}_r\left(\boldsymbol{a}_{\geq t}, \boldsymbol{w}_t \boldsymbol{a}_{\geq t+1}\right)$$

- *shook it*
- *kissed it fervently*
- *he tried not to shudder in relief*

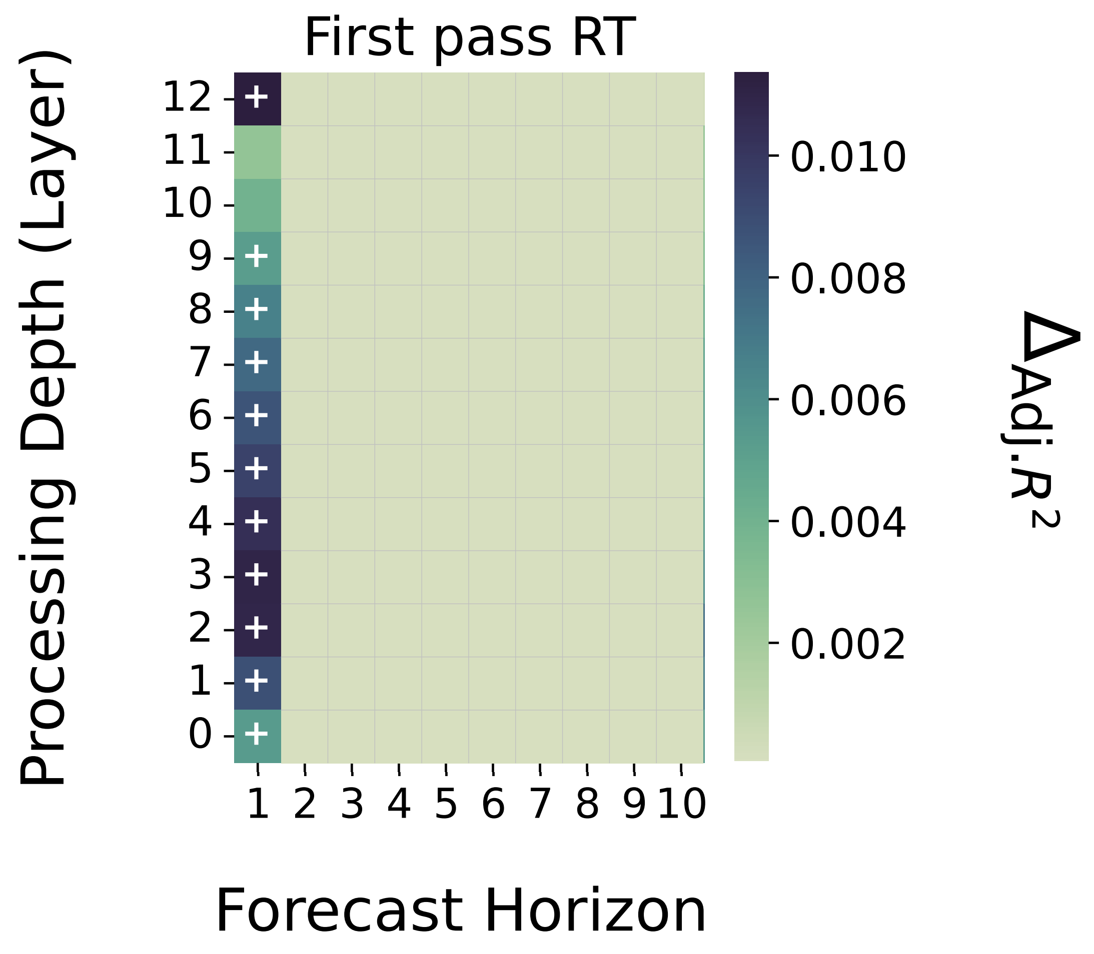
d_r

- *it fervently*
- *him tenderly on his forehead*
- *his palm*

Giulianelli, Wallbridge, Cotterell, Fernández. *Journal of Memory and Language*. 2026.

Incremental alternative sampling

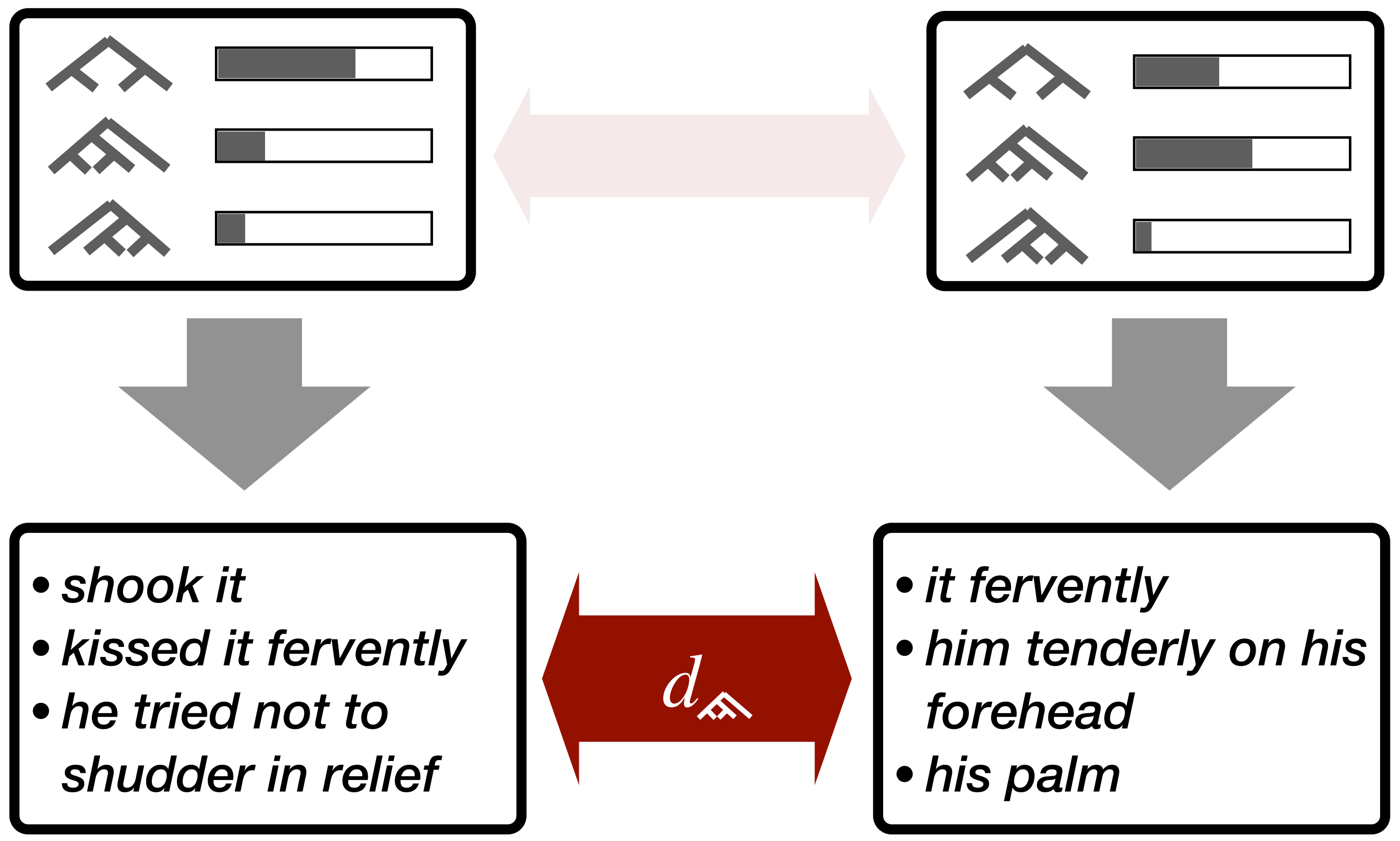
Temporal resolution: forecast horizon of 1…10 words
Representational resolution: LM layers from 0-th to last



Language model: GPT-2 Small
Stimuli: M = 1726 target–context pairs from English novels (de Varda et al. 2023)
Response: First-pass reading time (gaze duration)

Giulianelli, Wallbridge, Cotterell, Fernández. *Journal of Memory and Language*. 2026.

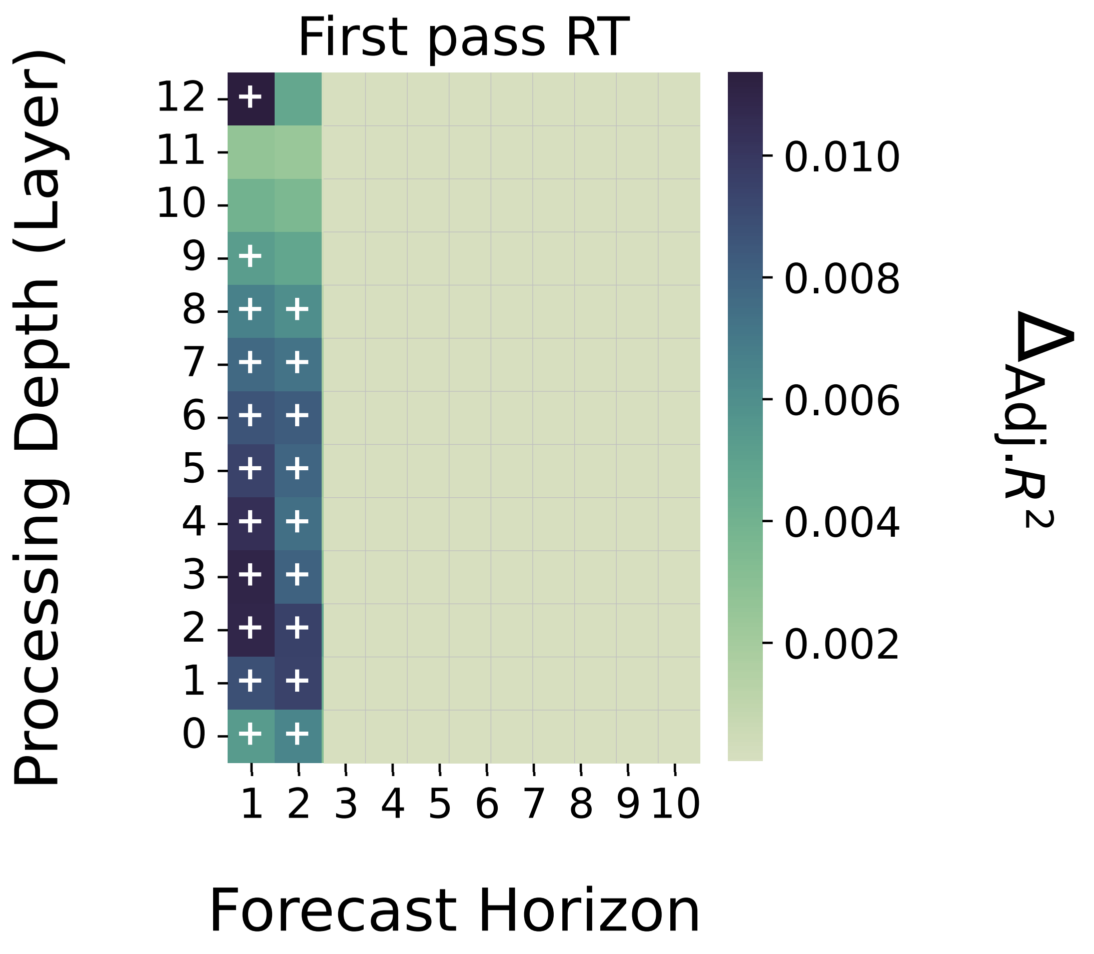
$$\mathbb{E}_{\boldsymbol{a}_t \sim p(\cdot \mid \boldsymbol{w}_{<t})} \mathbb{E}_{\boldsymbol{a}_{t+1} \sim p(\cdot \mid \boldsymbol{w}_{<t})} \, d\left(\boldsymbol{a}_t, w_t \boldsymbol{a}_{t+1}\right)$$



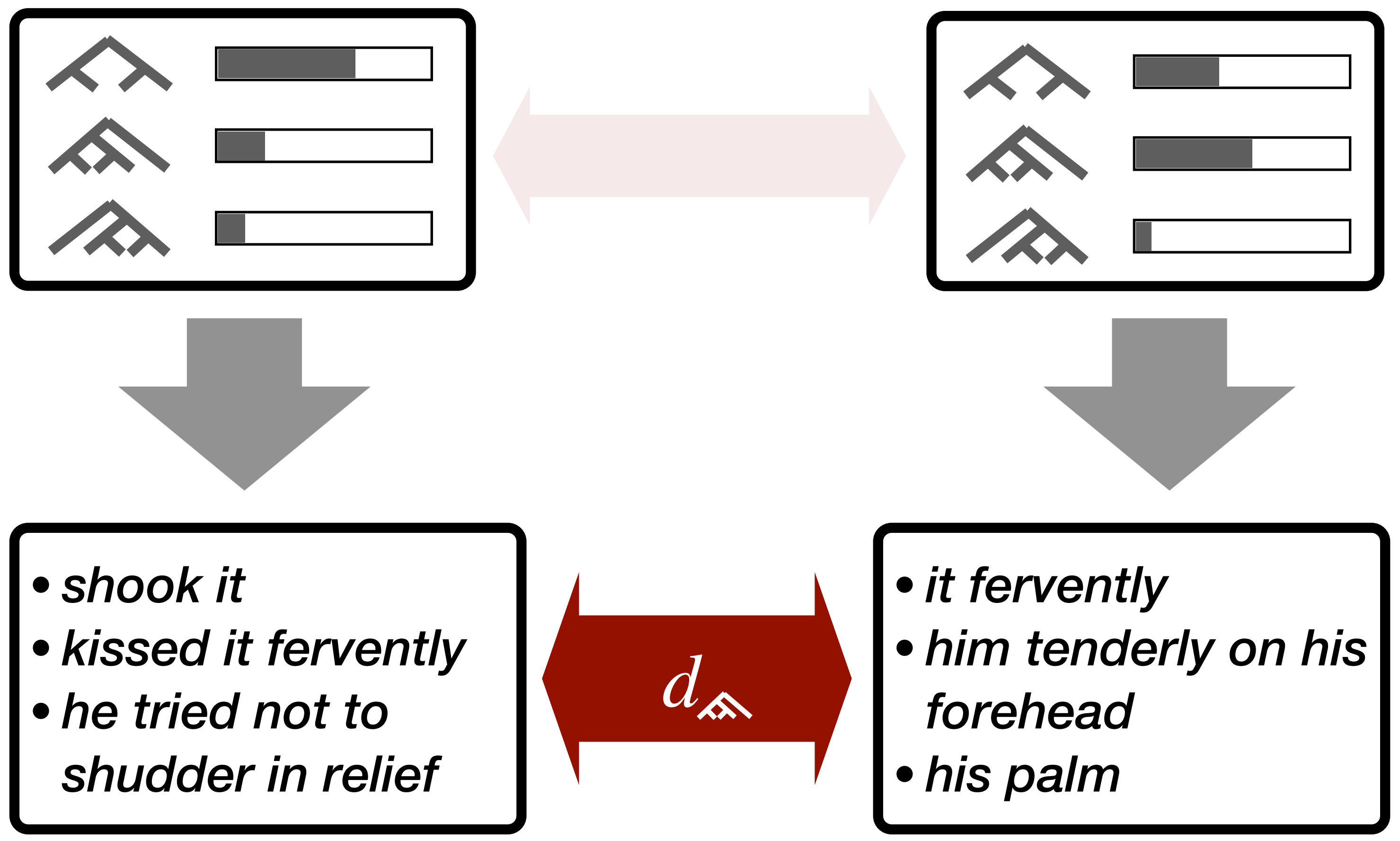
Allows explicitly manipulating expectations'
- **temporal** resolution (*forecast horizon)*
- **representational** resolution (*processing depth)*

Incremental alternative sampling

<u>Temporal resolution</u>: forecast horizon of 1…10 words
<u>Representational resolution</u>: LM layers from 0-th to last



$$\mathbb{E}_{\boldsymbol{a}_t\sim p(\cdot|\boldsymbol{w}_{<t})}\mathbb{E}_{\boldsymbol{a}_{t+1}\sim p(\cdot|\boldsymbol{w}_{<t})}\, d\left(\boldsymbol{a}_t, w_t\boldsymbol{a}_{t+1}\right)$$



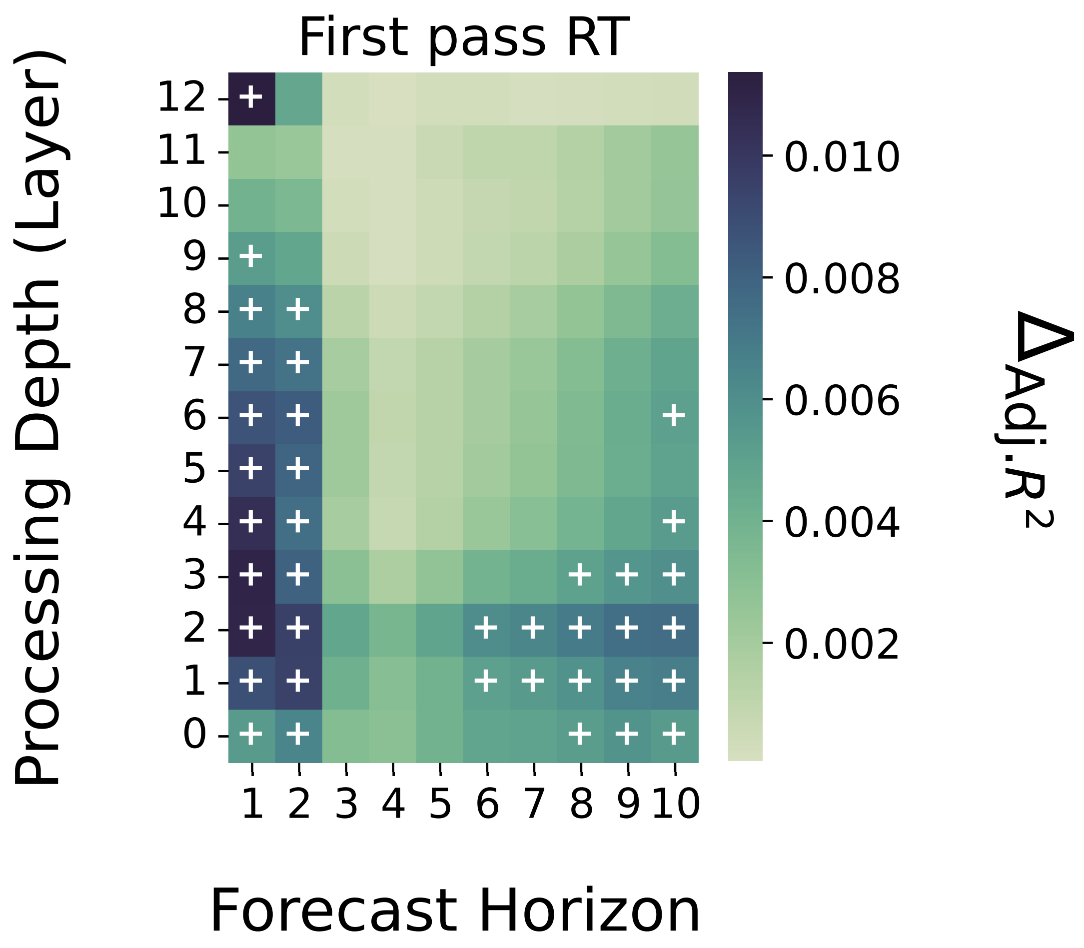
Allows explicitly manipulating expectations'
- **temporal** resolution (*forecast horizon)*
- **representational** resolution (*processing depth)*

<u>Language model</u>: GPT-2 Small
<u>Stimuli</u>: M = 1726 target–context pairs from English novels (de Varda et al. 2023)
<u>Response</u>: First-pass reading time (gaze duration)

Giulianelli, Wallbridge, Cotterell, Fernández. *Journal of Memory and Language.* 2026.

Incremental alternative sampling

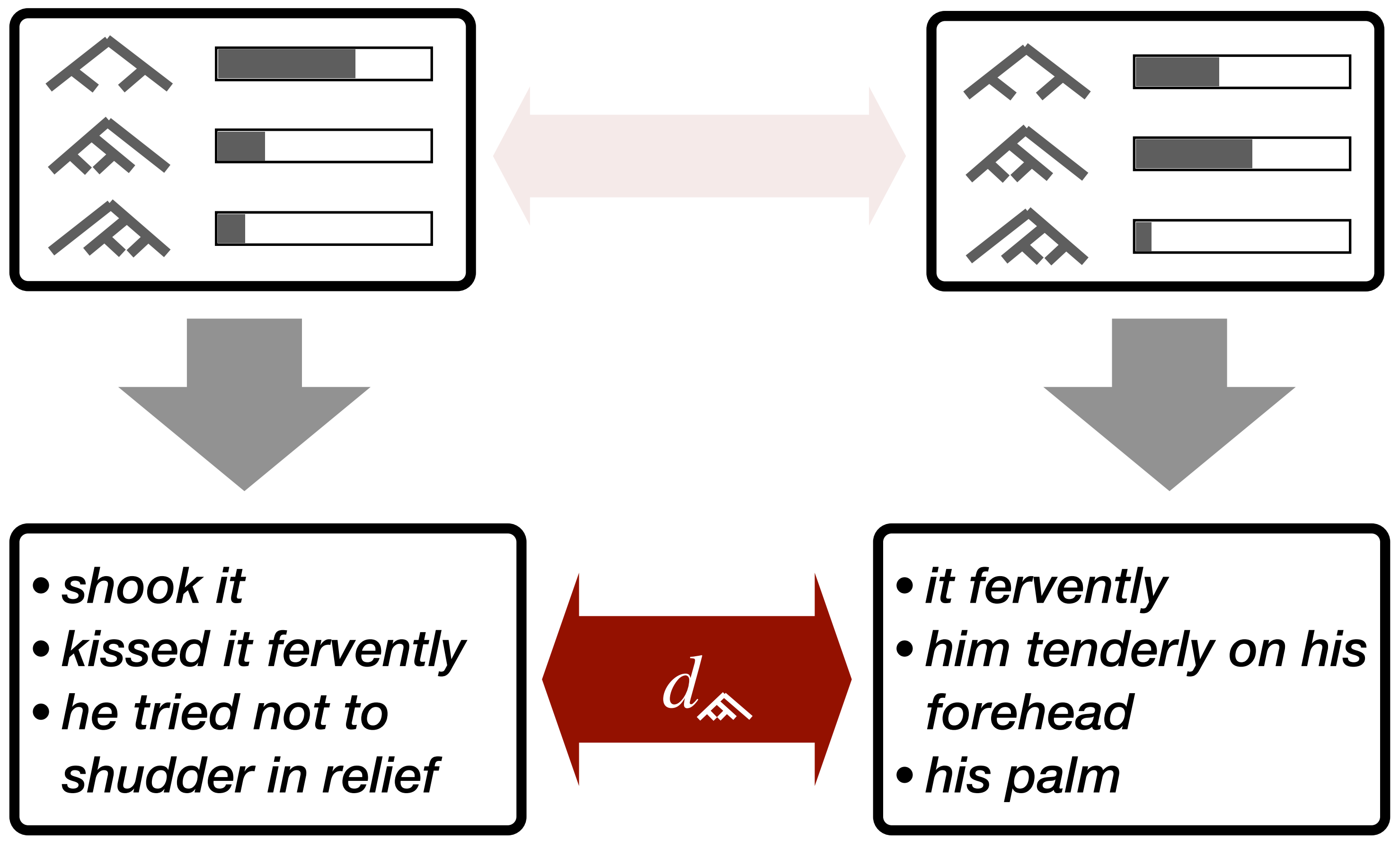
Temporal resolution: forecast horizon of 1…10 words
Representational resolution: LM layers from 0-th to last



Language model: GPT-2 Small
Stimuli: M = 1726 target–context pairs from English novels (de Varda et al. 2023)
Response: First-pass reading time (gaze duration)

Giulianelli, Wallbridge, Cotterell, Fernández. *Journal of Memory and Language.* 2026.

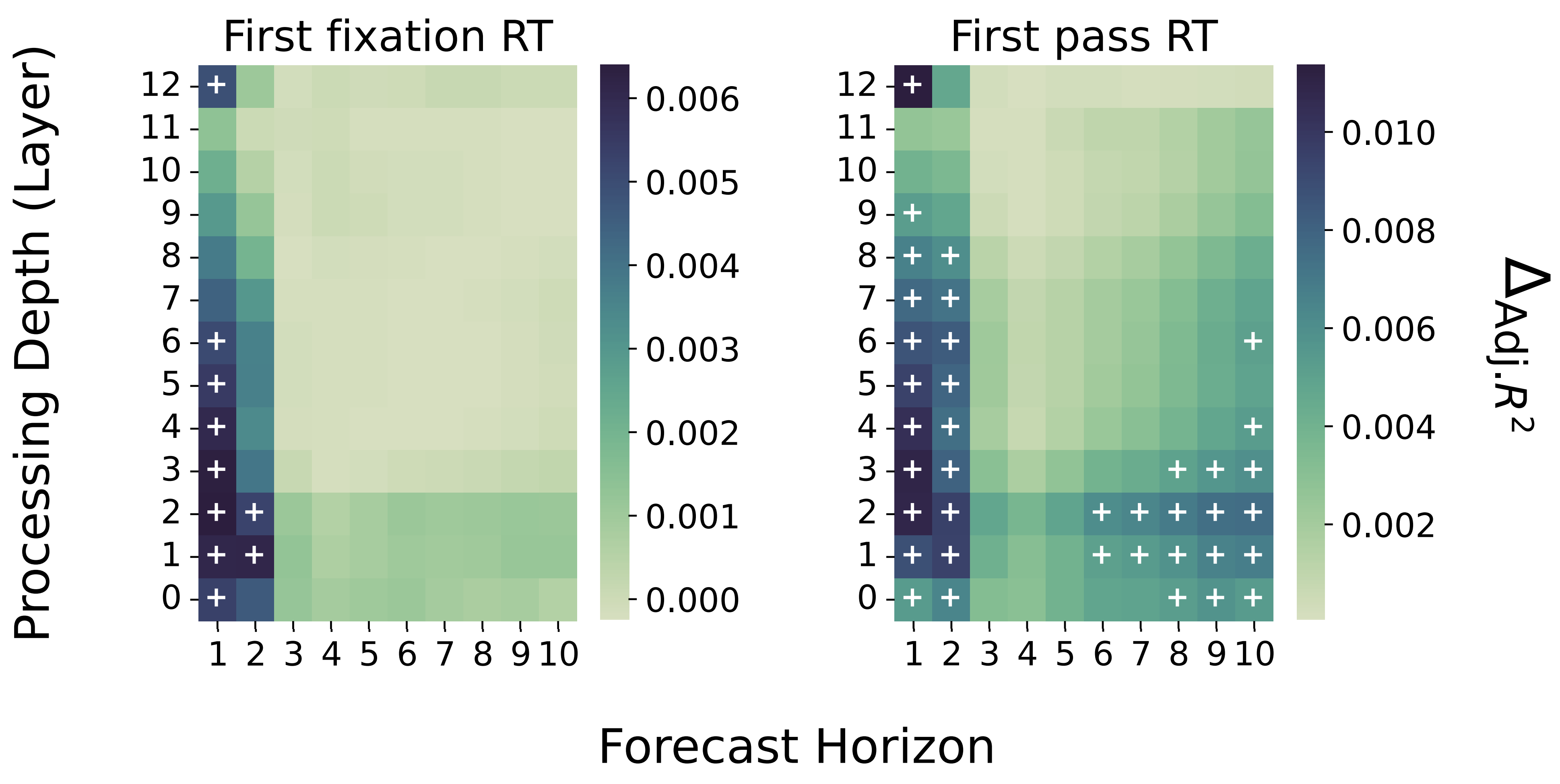
$$\mathbb{E}_{\boldsymbol{a}_t \sim p(\cdot|\boldsymbol{w}_{<t})} \mathbb{E}_{\boldsymbol{a}_{t+1} \sim p(\cdot|\boldsymbol{w}_{<t})} \; d\left(\boldsymbol{a}_t, w_t \boldsymbol{a}_{t+1}\right)$$



Allows explicitly manipulating expectations'
- **temporal** resolution (*forecast horizon)*
- **representational** resolution (*processing depth)*

Incremental alternative sampling

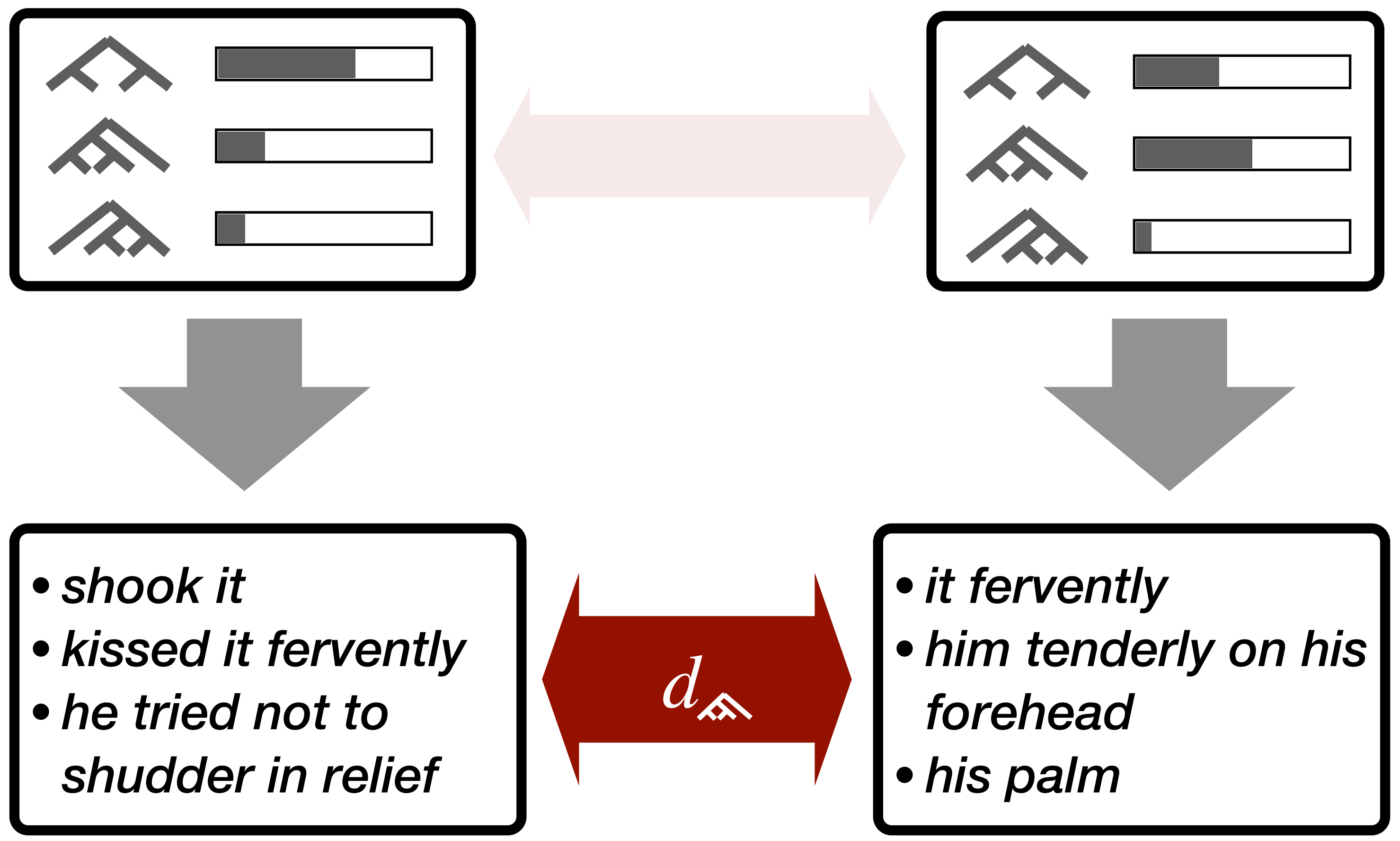
Temporal resolution: forecast horizon of 1…10 words
Representational resolution: LM layers from 0-th to last



Language model: GPT-2 Small
Stimuli: M = 1726 target–context pairs from English novels (de Varda et al. 2023)
Response: First-fixation and first-pass reading time; N400

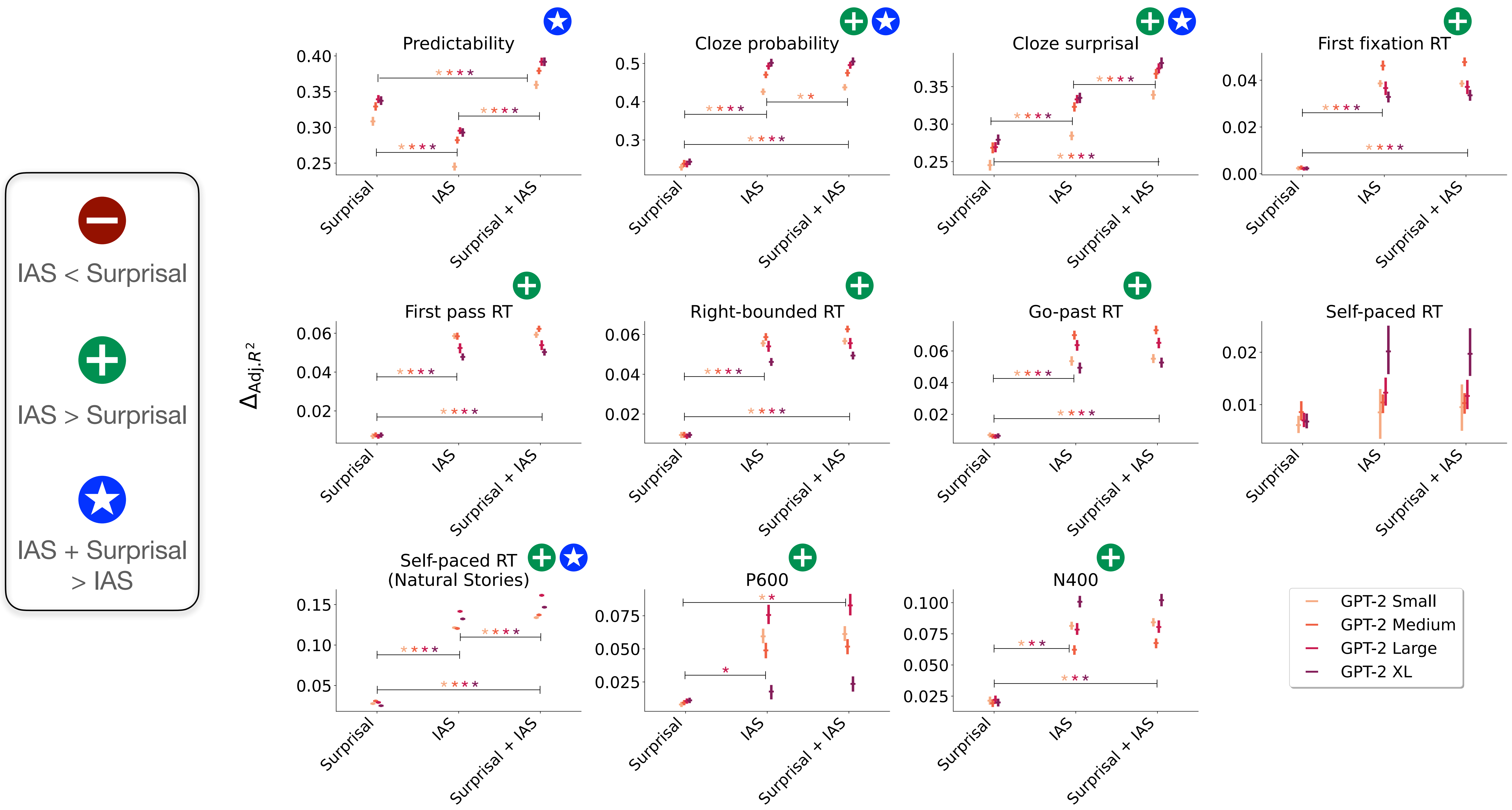
Giulianelli, Wallbridge, Cotterell, Fernández. *Journal of Memory and Language.* 2026.

$$\mathbb{E}_{\boldsymbol{a}_t \sim p(\cdot|\boldsymbol{w}_{<t})} \mathbb{E}_{\boldsymbol{a}_{t+1} \sim p(\cdot|\boldsymbol{w}_{<t})} \; d\left(\boldsymbol{a}_t, w_t \boldsymbol{a}_{t+1}\right)$$



Allows explicitly manipulating expectations'
- **temporal** resolution (*forecast horizon)*
- **representational** resolution (*processing depth)*

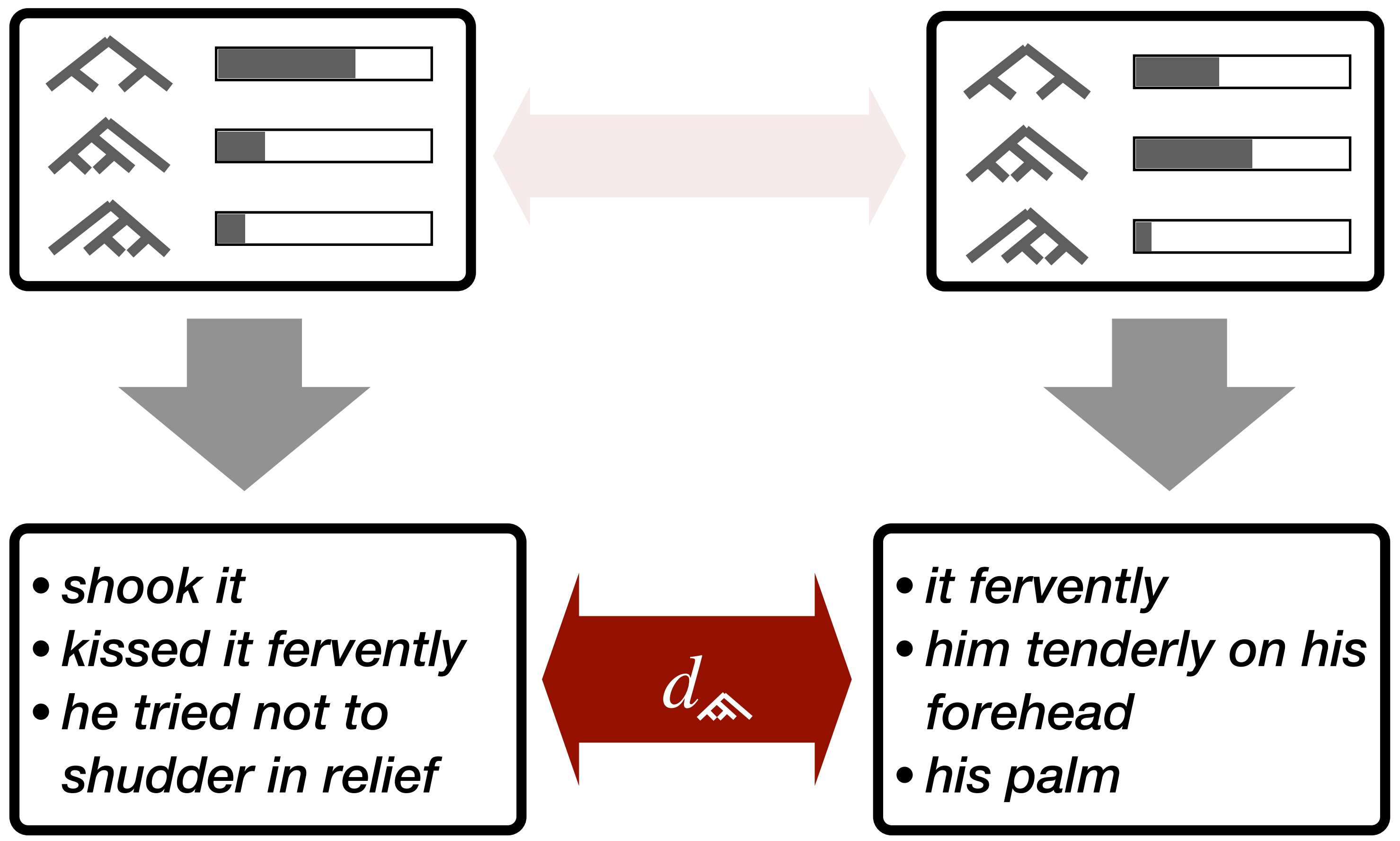
Incremental alternative sampling



Language model: GPT-2 Small, Medium, Large, XL
Stimuli: M = 1726 target–context pairs from English novels (de Varda et al. 2023)
Response: predictability ratings, cloze (log) probability, ERPs, reading times

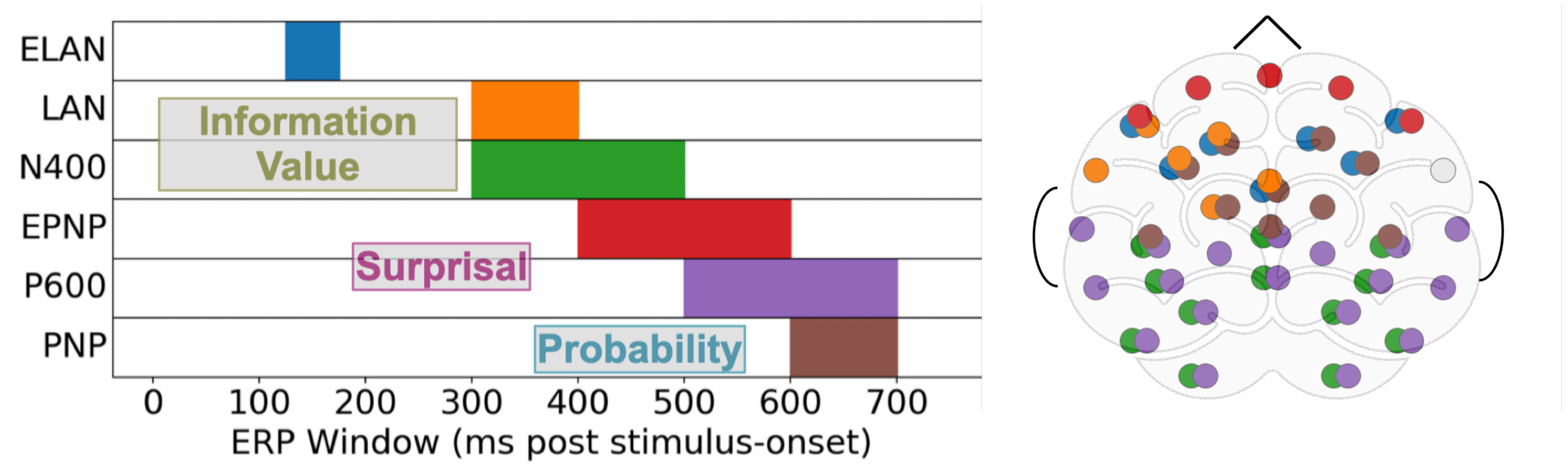
Giulianelli, Wallbridge, Cotterell, Fernández. *Journal of Memory and Language.* 2026.

$$\mathbb{E}_{\boldsymbol{a}_t\sim p(\cdot|\boldsymbol{w}_{<t})}\mathbb{E}_{\boldsymbol{a}_{t+1}\sim p(\cdot|\boldsymbol{w}_{<t})}\; d\left(\boldsymbol{a}_t, w_t\boldsymbol{a}_{t+1}\right)$$



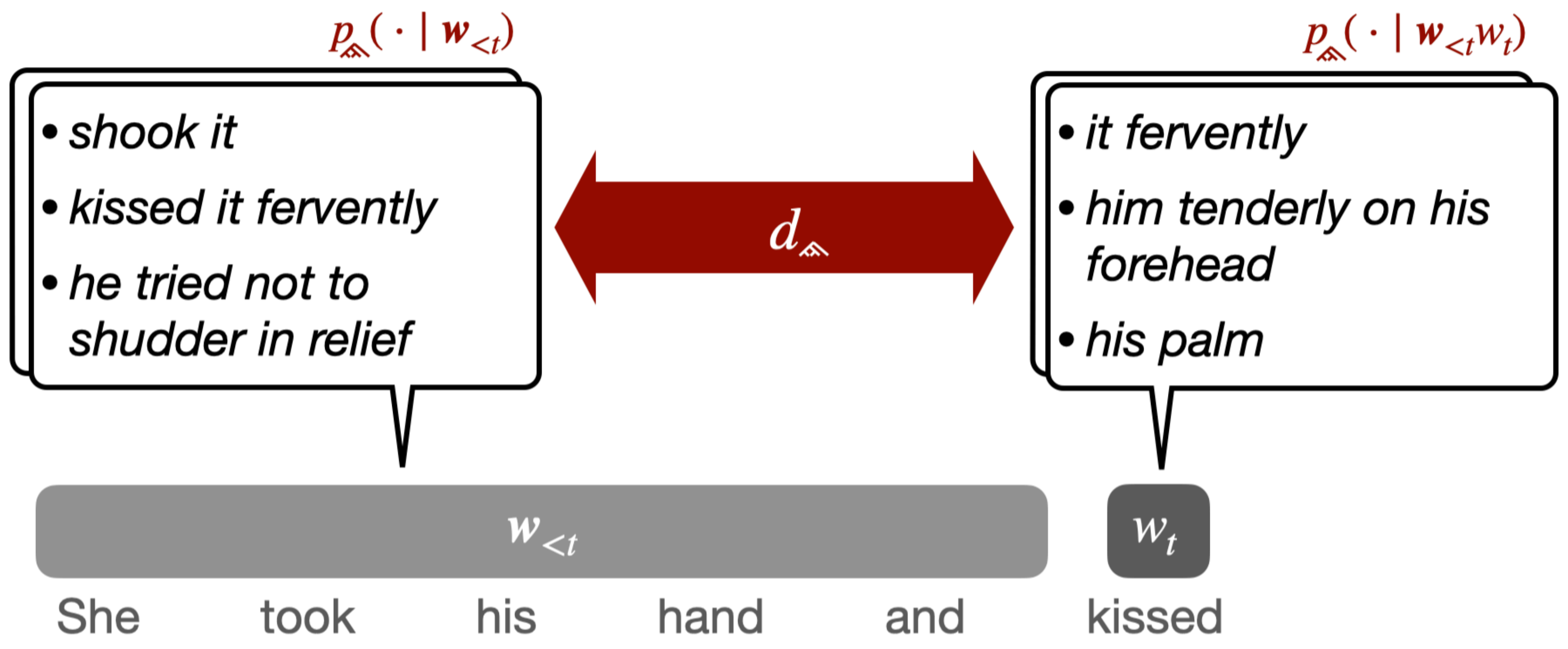
Allows explicitly manipulating expectations'
- **temporal** resolution (*forecast horizon)*
- **representational** resolution (*processing depth)*

What's next?



Targeted studies of phenomena like garden path effects

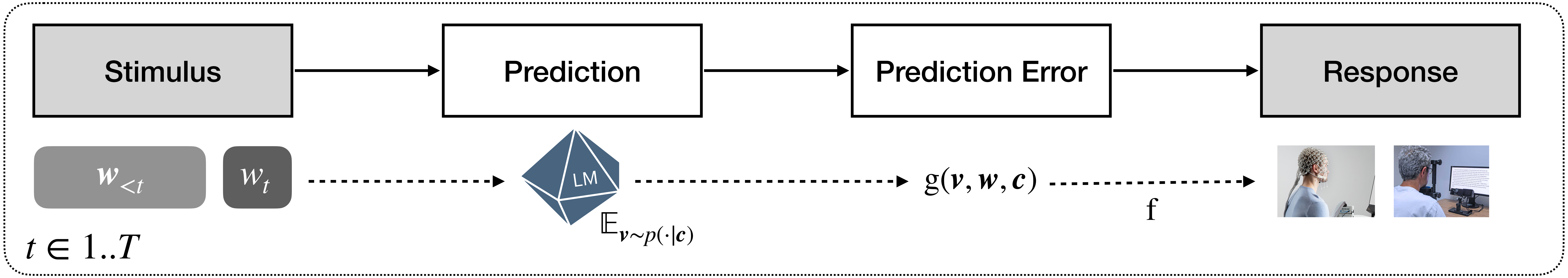
Spatio-temporal modelling of continuous EEG signals



Targeted representation functions (better syntactic ones)

Anticipatory uncertainty

Reasoning over alternatives



Beyond the surprisal model of incremental processing difficulty

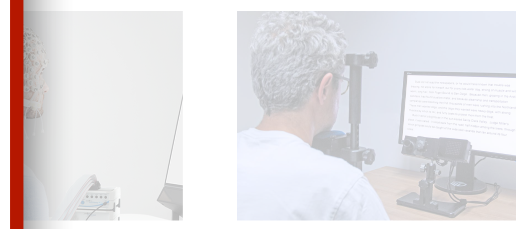
LM training and evaluation using alternative information-theoretic metrics

Destructuring Surprisal Theory

BBC wants to

Stimulus

Response



$t \in 1..T$

Theoretical issues

- ✦ conflates levels of linguistic processing
  Giulianelli, Wallbridge, Fernández. EMNLP 2023;
  Meister, Giulianelli, Pimentel. EMNLP 2024.
- ✦ conflates expectations at varying temporal resolutions
  Giulianelli, Wallbridge, Cotterell, Fernández. JML 2026.
- ✦ classical derivation requires strong assumptions
  Giulianelli, Baan, Aziz, Fernández, Plank. EMNLP 2023.
  Giulianelli, Wallbridge, Cotterell, Fernández. JML 2026.
- ✦ considers only responsive uncertainty
  Giulianelli, Opedal, Cotterell. Findings of EMNLP 2024.
- ✦ is a special case of a more general information-theoretic model
  Giulianelli, Opedal, Cotterell. Findings of EMNLP 2024.

Issues of the methodological paradigm

- ✦ word- or character-level stimuli vs. token-level LMs
  Giulianelli, Malagutti, Gastaldi, DuSell, Vieira, Cotterell. EMNLP 2024.
  Vieira, LeBrun, Giulianelli, Gastaldi, DuSell, Terilla, O'Donnell, Cotterell. ICML 2025.
- ✦ word-level aggregations of continuous data
  Re, Opedal, Manaiev, Giulianelli, Cotterell. EMNLP 2025.

Token-level LMs for character-level problems

`Anne␣lost␣control␣and␣laughed.`

Stimulus

Skip Rate (`control␣`|`Anne␣lost␣`)

Measurement

$-\log p($`con`|`Anne␣lost␣`$)$

Predictor

Vieira, LeBrun, Giulianelli, Gastaldi, DuSell, Terilla, O'Donnel, Cotterell. ICML 2025.

Full ROI	⟨`lost␣`, `control␣`, `and␣`, `laughed.`⟩
Fixed	⟨`los`, `con`, `and`, `lau`⟩
Dynamic (7)	⟨`lost␣`, `contr`, `and`, `laugh`⟩
Dynamic (8)	⟨`lost␣`, `contro`, `and␣`, `laughe`⟩

Giulianelli, Malagutti, Gastaldi, DuSell, Vieira, Cotterell. EMNLP 2024.

Token-level LMs for character-level problems

`Anne␣lost␣control␣and␣laughed.`

Stimulus

Skip Rate (`control␣`|`Anne␣lost␣`)

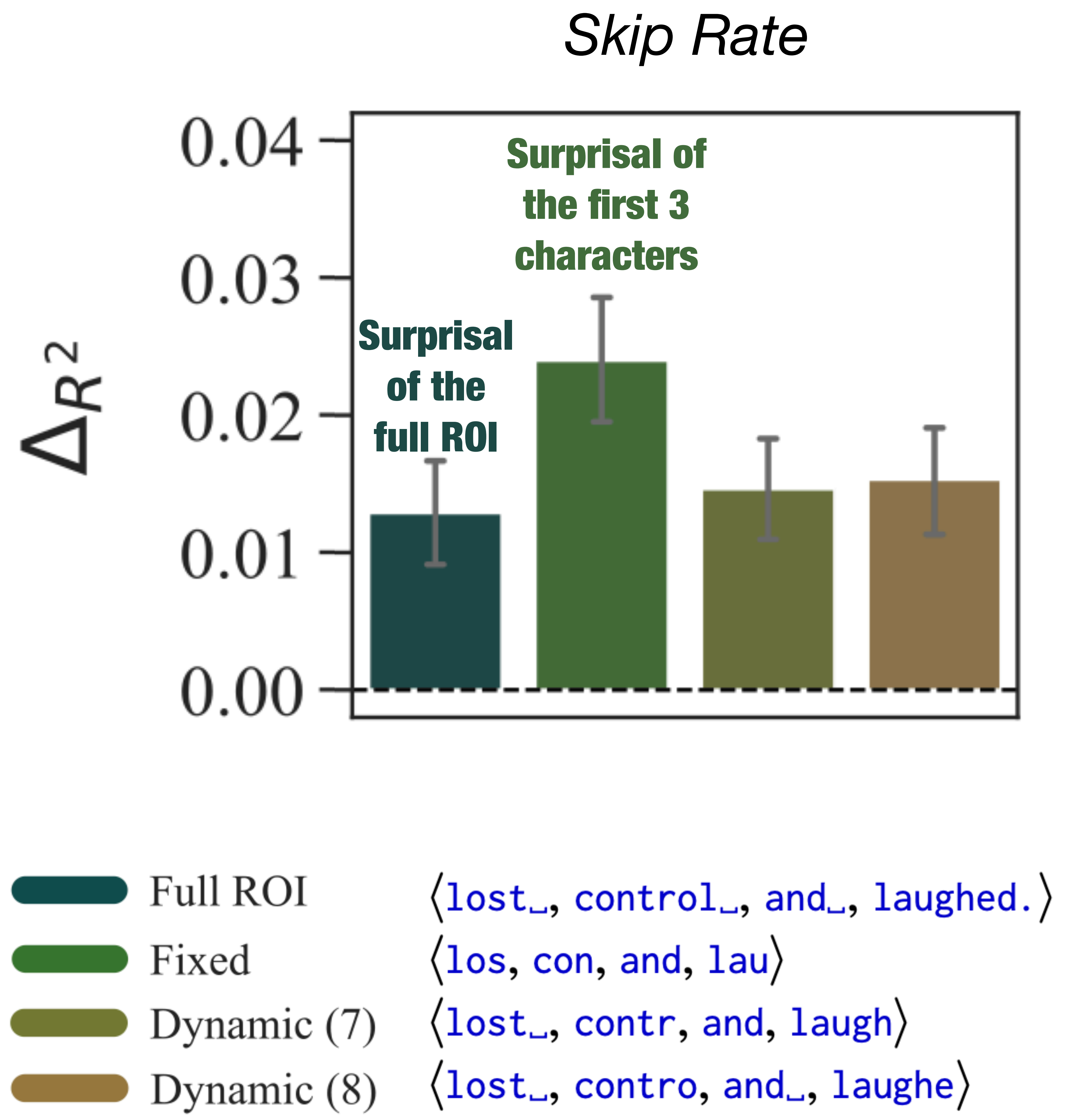
Measurement

$-\log p(\texttt{con}|\texttt{Anne␣lost␣})$

Predictor

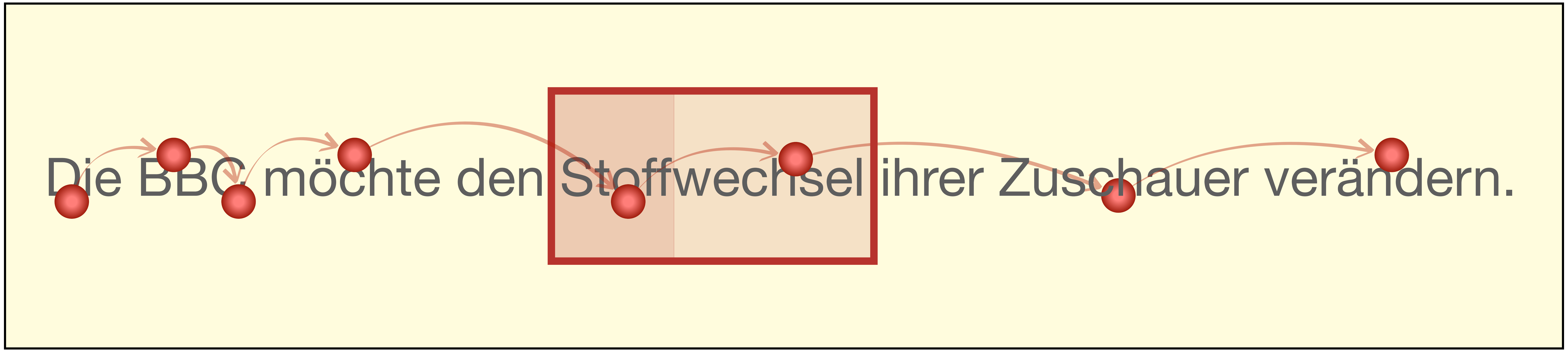
Token-level LM → *realphabetization* → Character-level LM

Vieira, LeBrun, Giulianelli, Gastaldi, DuSell, Terilla, O'Donnel, Cotterell. ICML 2025.



Giulianelli, Malagutti, Gastaldi, DuSell, Vieira, Cotterell. EMNLP 2024.

Lossy treatment of spatio-temporal data

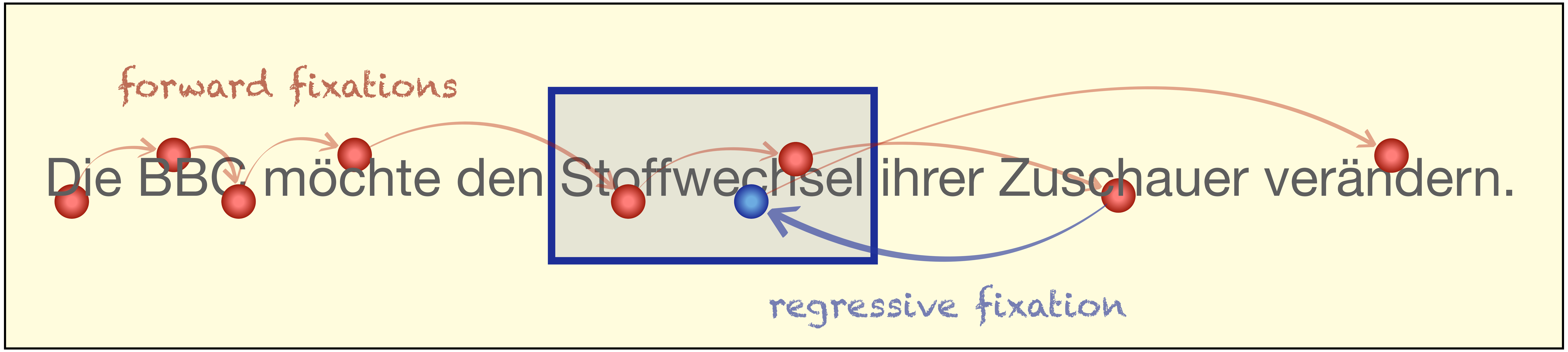


Words assumed as the fundamental unit of incremental processing

→ Language models defined over arbitrary vocabulary of “**tokens**”

→ Ignores **other plausible regions of interest** *(e.g., characters or morphemes)*

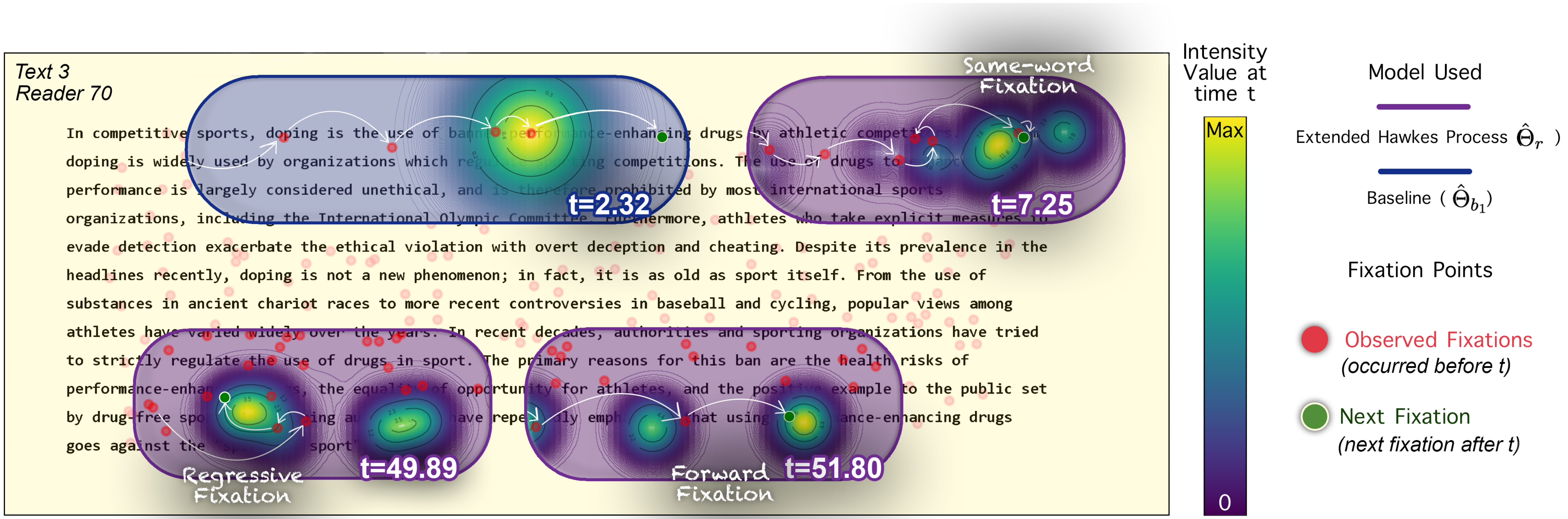
Lossy treatment of spatio-temporal data



Multiple **raw data points** *(e.g., individual fixations)* **aggregated** into a single measurement *(e.g., total fixation time)*

➡ Obscures **distinct underlying cognitive processes**

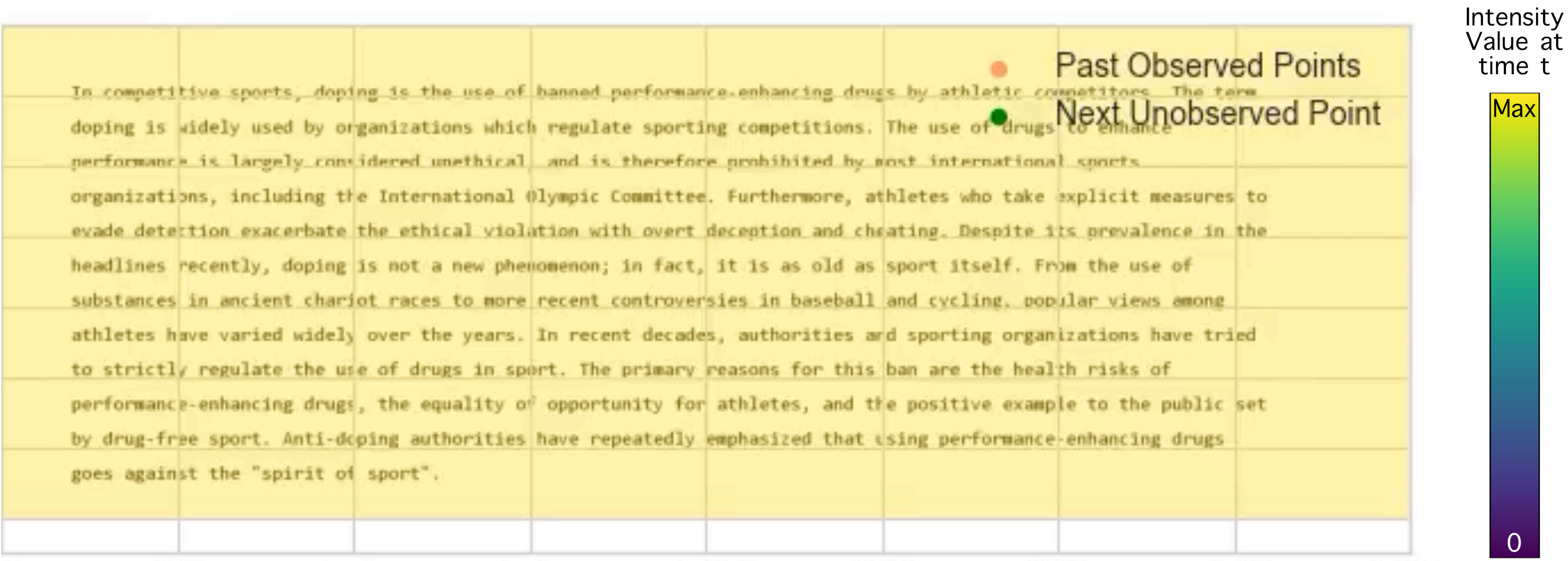
Fine-grained spatio-temporal modeling of reading behaviour



Predicted Gaze Intensity map of a **spatio-temporal Hawkes process** with type-writer effect, previous fixation surprisal, and reader-specific coefficients.

Re, Opedal, Manaiev, Giulianelli, Cotterell. *ACL 2025*.

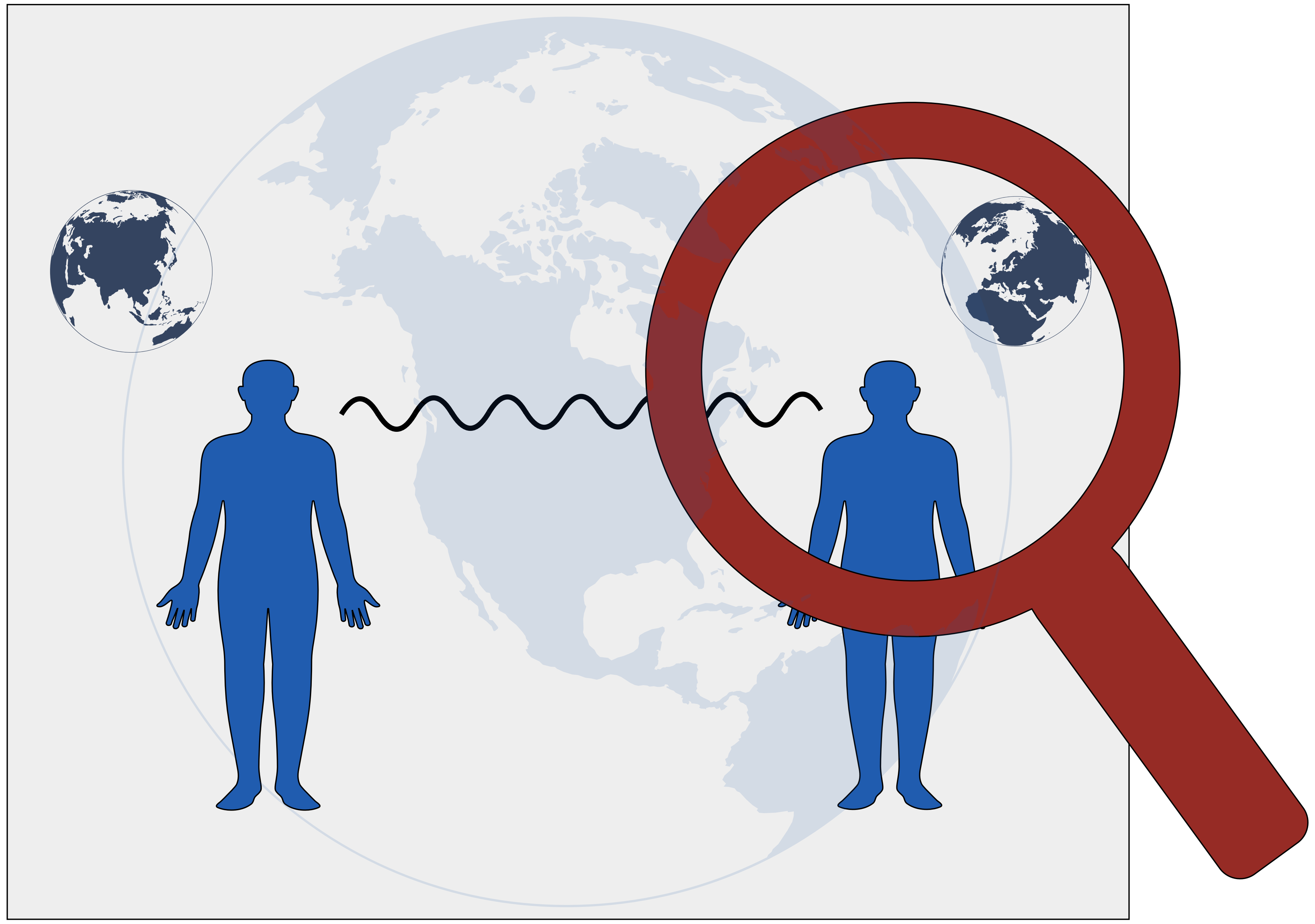
Fine-grained spatio-temporal modeling of reading behaviour



Predicted Gaze Intensity map of a **spatio-temporal Hawkes process** with type-writer effect, previous fixation surprisal, and reader-specific coefficients.

Re, Opedal, Manaiev, Giulianelli, Cotterell. *ACL 2025*.

Language comprehension

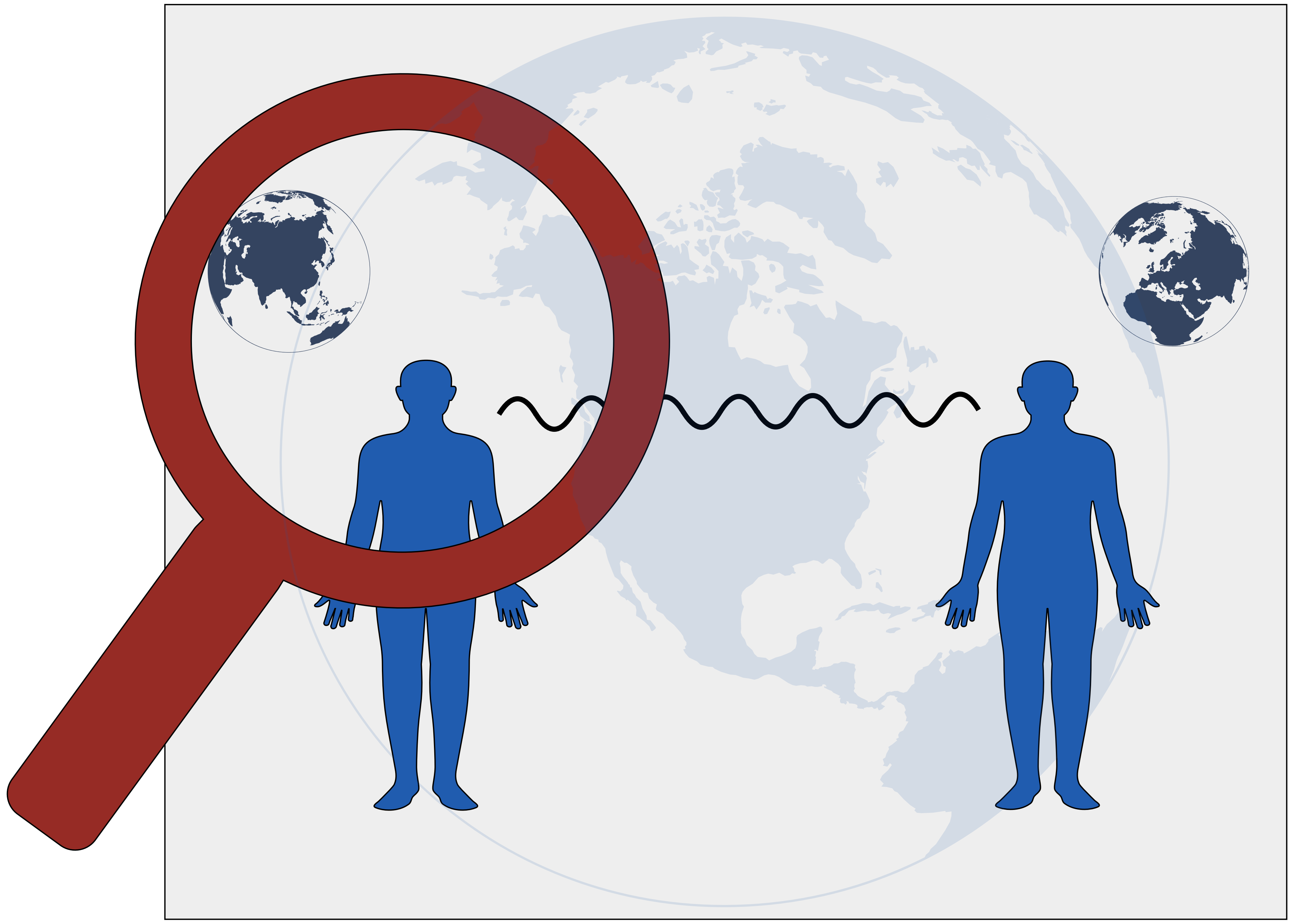


What is the role of prediction in language processing?

At which representational and temporal resolution does prediction take place?

Can behavioural and neural responses to language input be explained in terms of the input's information profile?

Language production

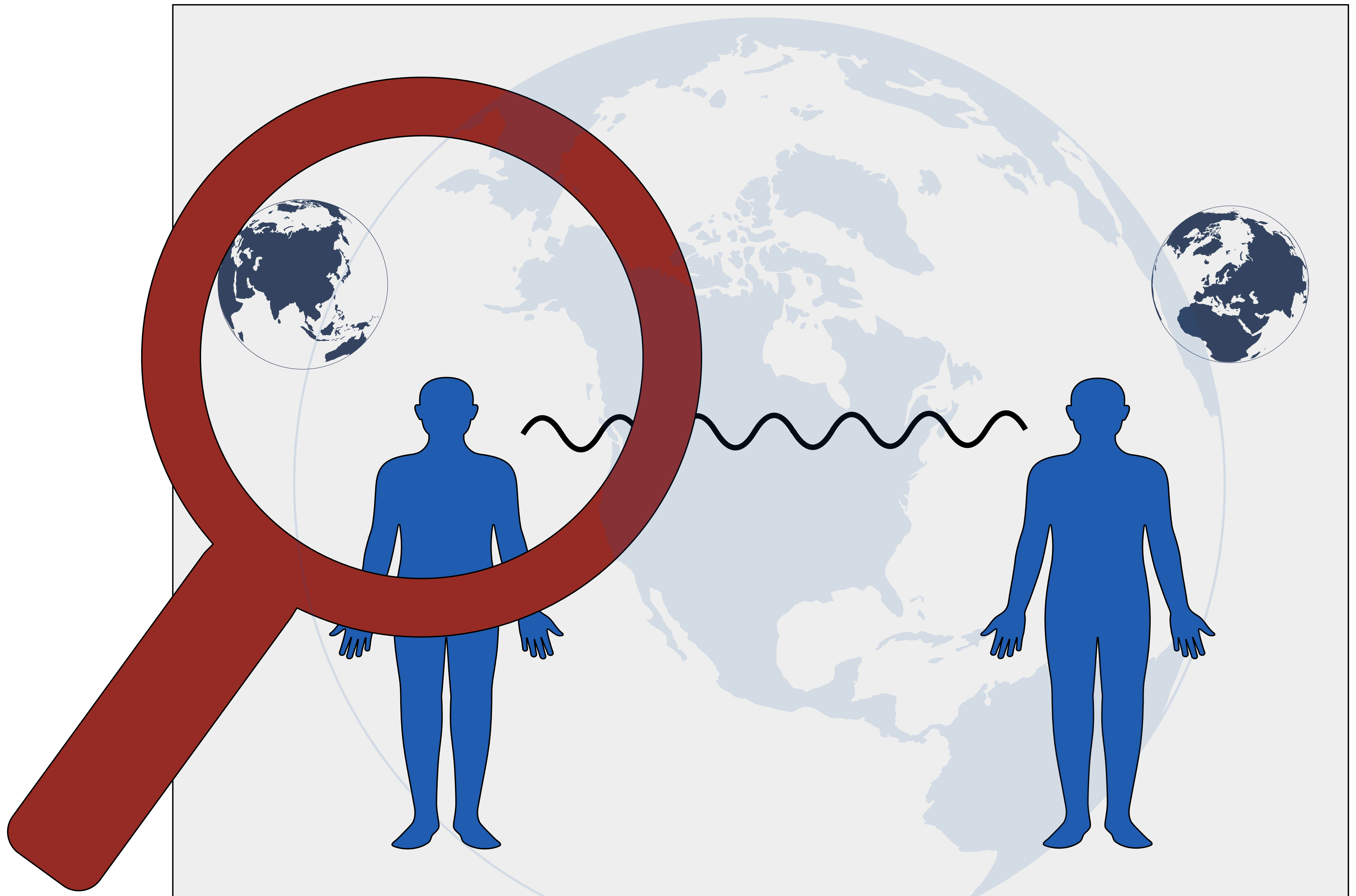


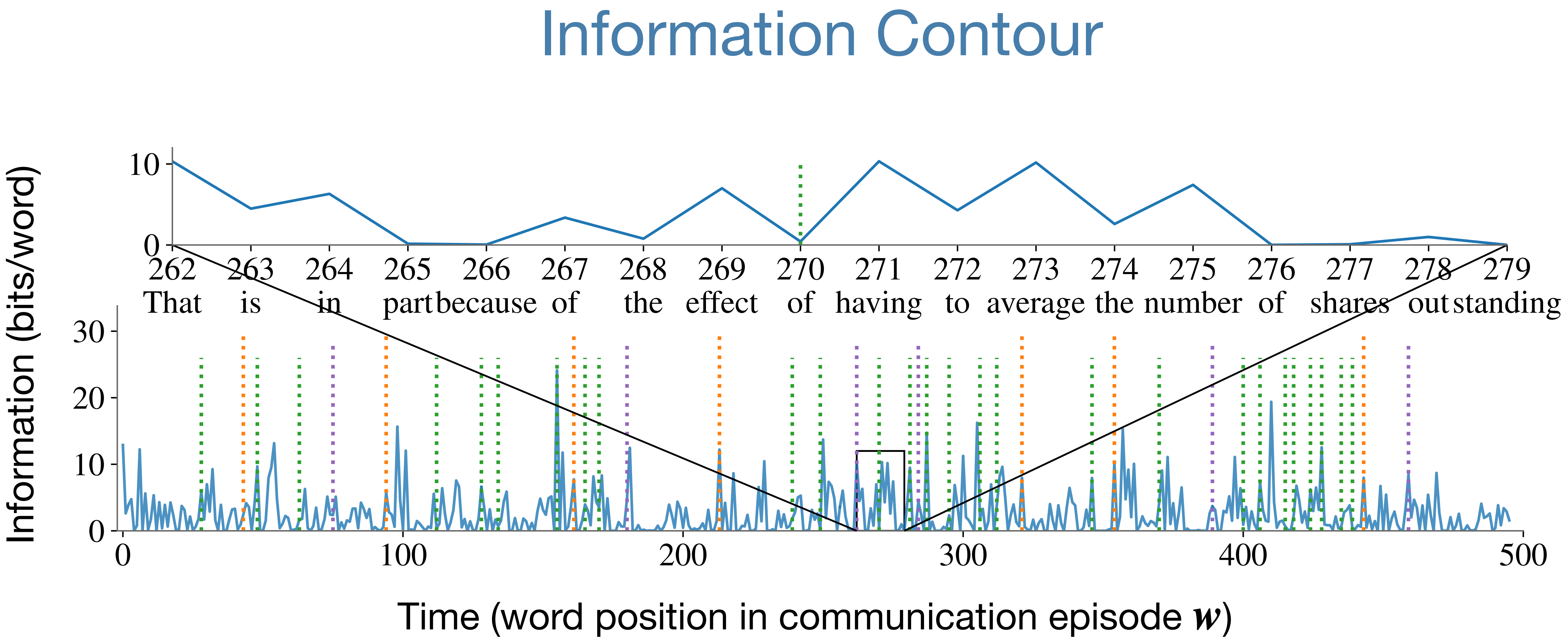
What is the rate at which producers transmit information?

Do producers make rational use of the communication channel?

How does context (linguistic and non-linguistic) modulate information rate?

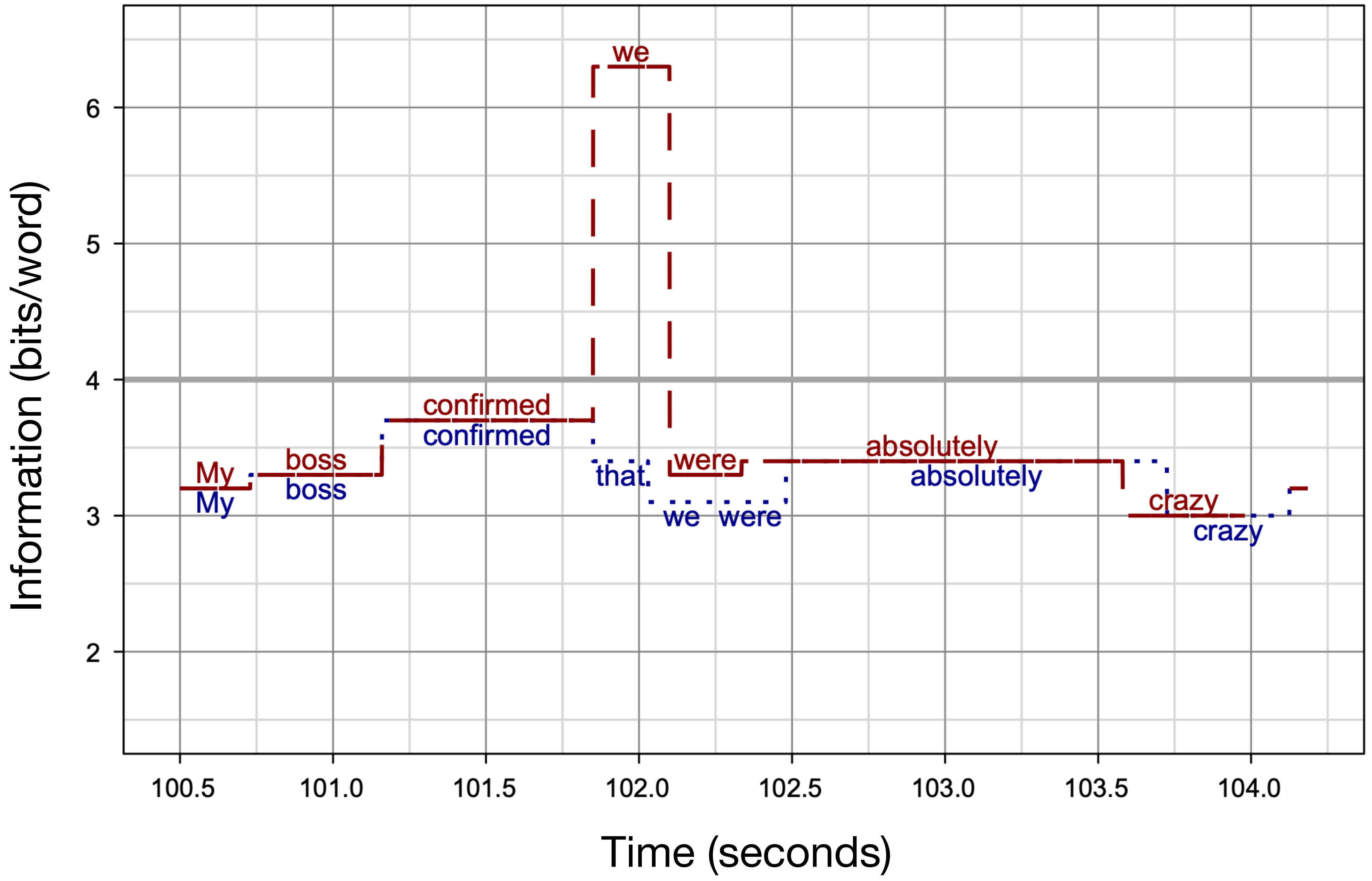
Information contours in texts and dialogues





$$-\log p(w_t \mid \boldsymbol{w}_{<t})$$

Information contours in texts and dialogues

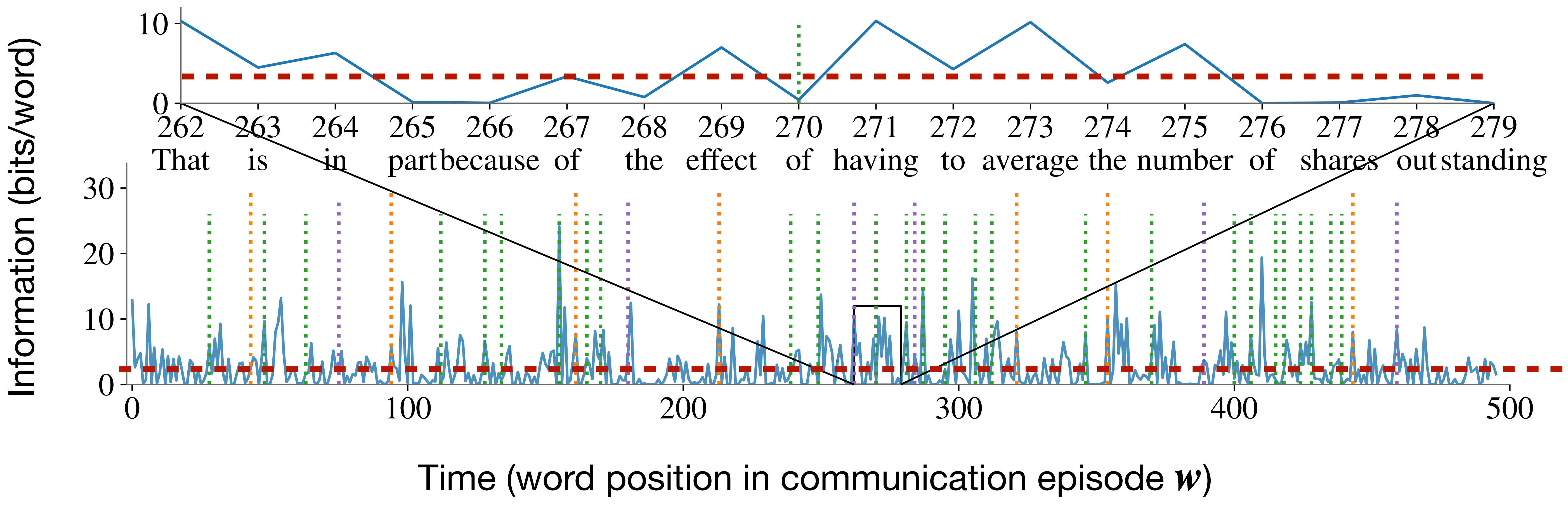


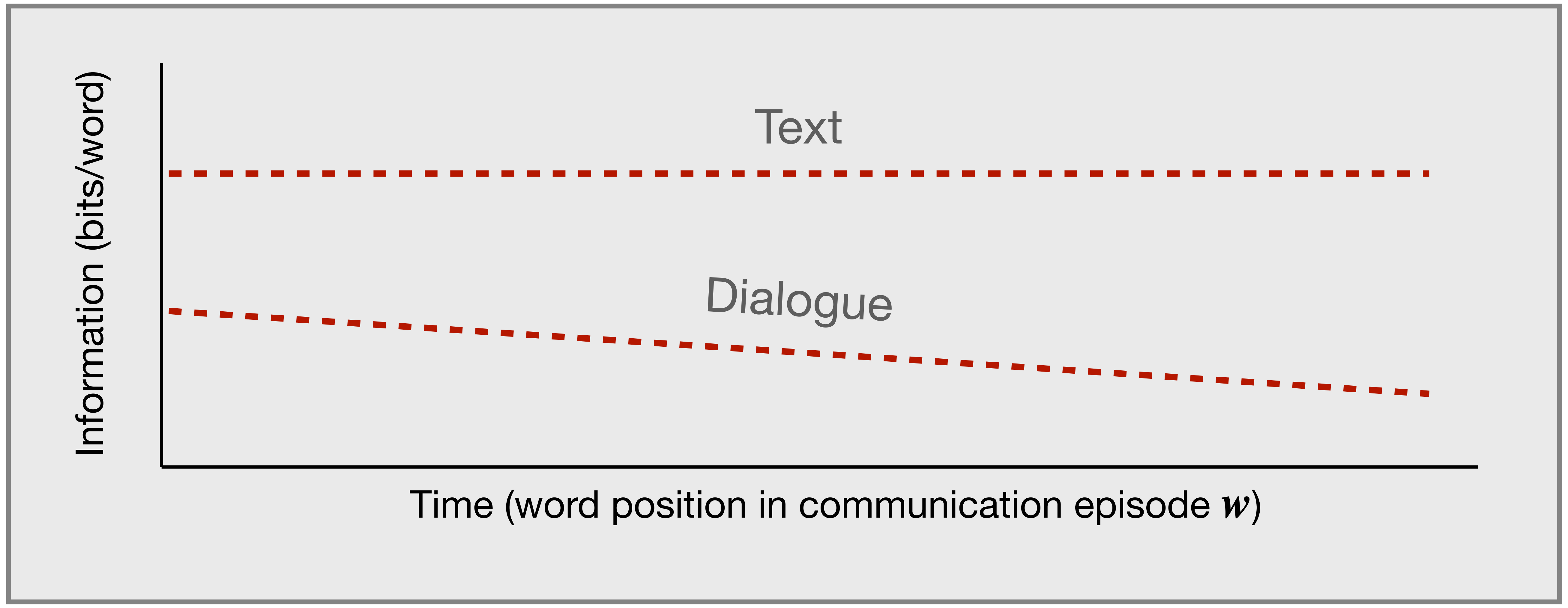
Jaeger. *Cognitive Psychology* (2010).

Hypothesis 1: Uniform Information Density

Subject to the constraints of the grammar, speakers optimise their linguistic signals such that the surprisals ι_w are distributed as uniformly as possible throughout a communication episode w.

Information contours in texts and dialogues





Hypothesis 1: Uniform Information Density

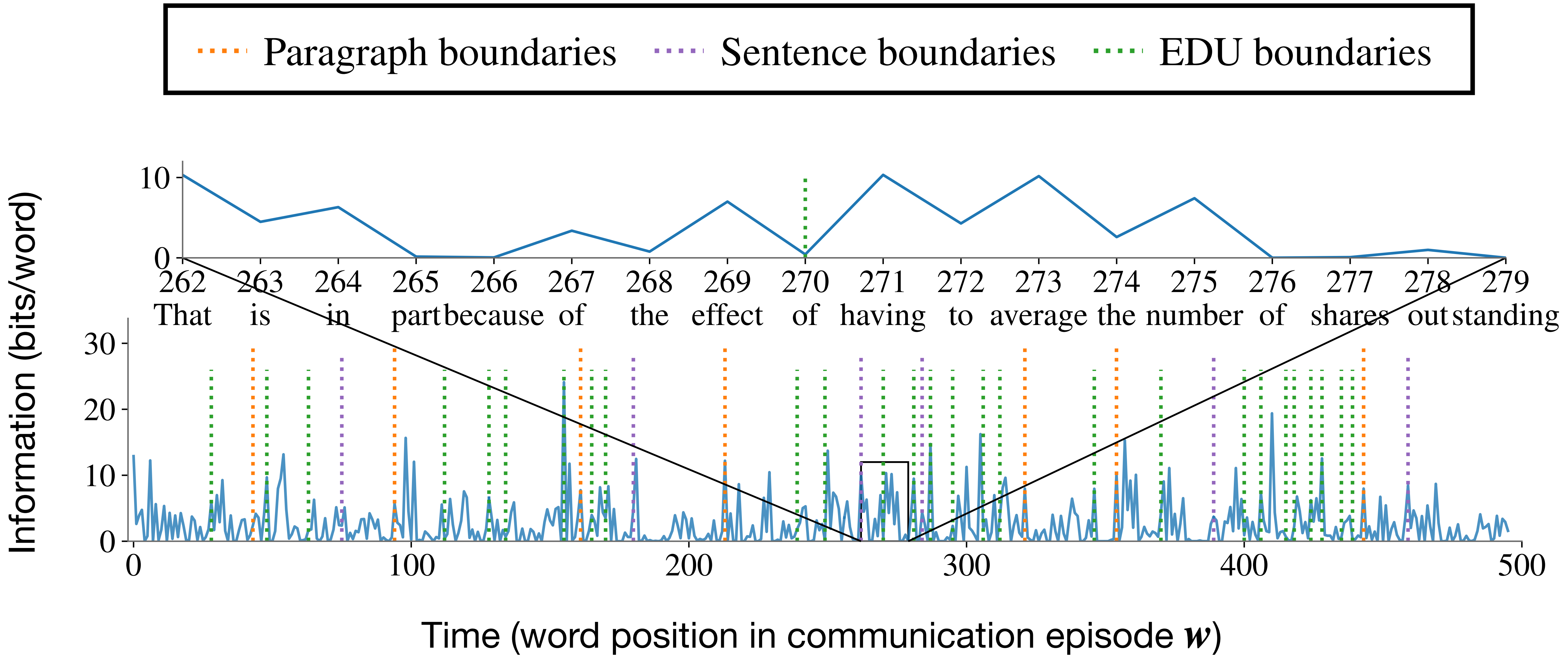
Subject to the constraints of the grammar, speakers optimise their linguistic signals such that the surprisals ι_w are distributed as uniformly as possible throughout a communication episode $\boldsymbol{w}$.

→ no evidence of local uniformity, pressure toward a global mean

→ information rate decreases in dialogues

Giulianelli & Fernández. CoNLL 2021.
Giulianelli, Sinclair, Fernández. AACL 2021.

Information contours in texts and dialogues



Hypothesis 1: Uniform Information Density

Subject to the constraints of the grammar, speakers optimise their linguistic signals such that the surprisals $\iota_{\boldsymbol{w}}$ are distributed as uniformly as possible throughout a communication episode $\boldsymbol{w}$.

→ no evidence of local uniformity, pressure toward a global mean

→ information rate decreases in dialogues

Giulianelli & Fernández. CoNLL 2021.
Giulianelli, Sinclair, Fernández. AACL 2021.

Hypothesis 2: Structured Context

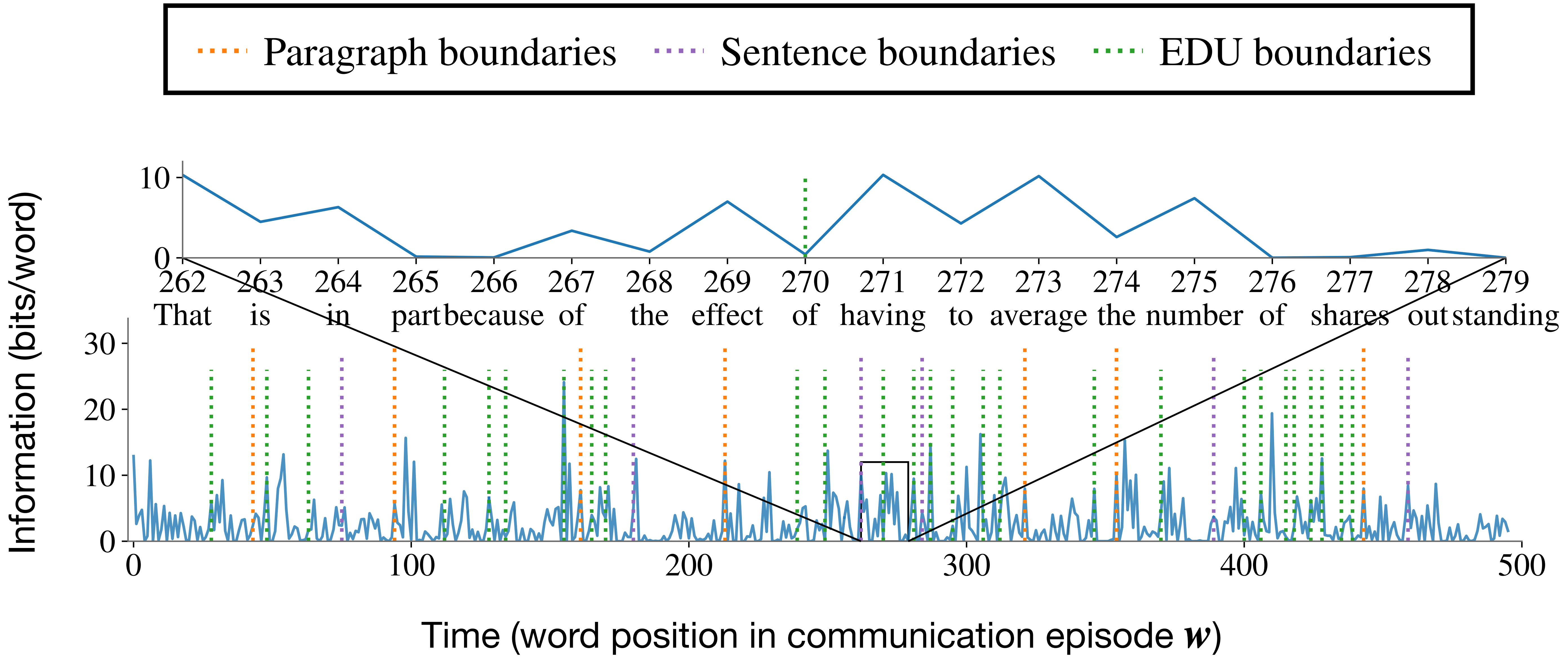
Values $\iota\left(w_t; \boldsymbol{w}_{<t}\right)$ in the surprisal contour $\iota_{\boldsymbol{w}}$ of a communication episode $\boldsymbol{w}$ are (partially) determined by the position of w_t within the hierarchy of $\boldsymbol{w}$'s constituent structural units.

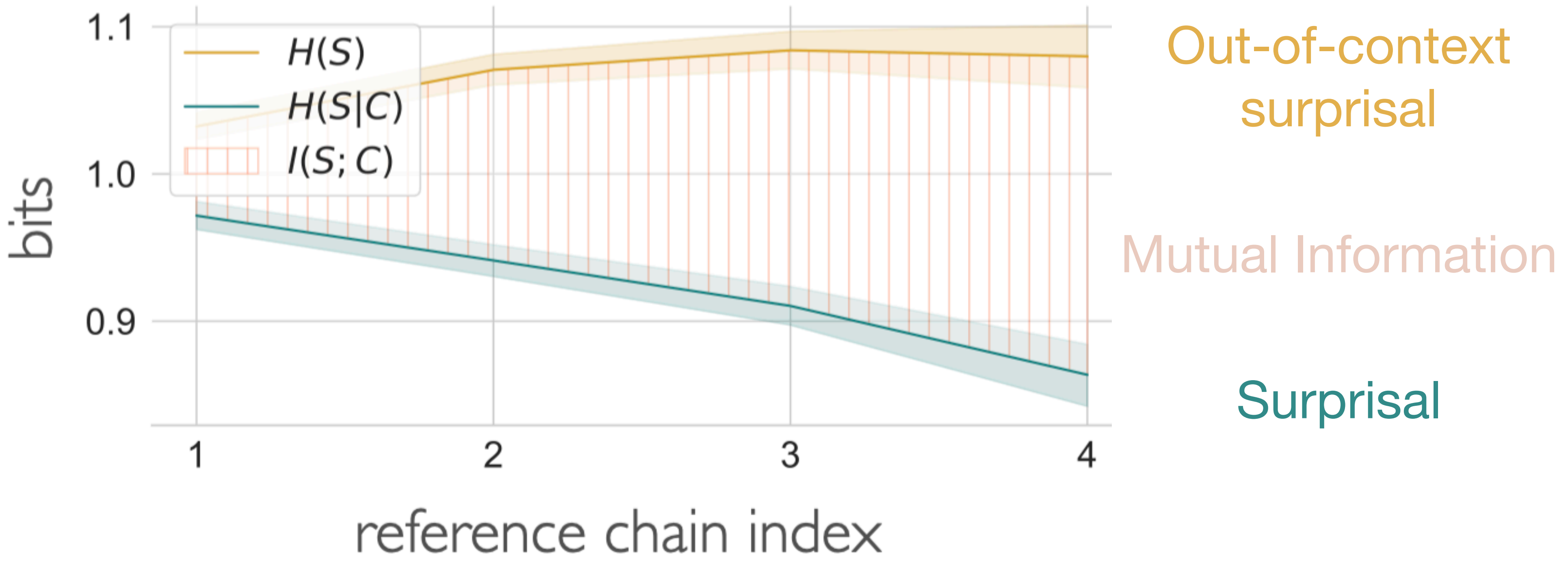
→ a unit's position within **contextual structure** predicts its surprisal

- → RST discourse units in texts
- → task-specific contextual units in dialogues

Giulianelli, Sinclair, Fernández. EMNLP 2021.
Tsipidi, Nowak, Cotterell, Wilcox, Giulianelli, Warstadt. EMNLP 2024.

Information contours in texts and dialogues





Hypothesis 1: Uniform Information Density

Subject to the constraints of the grammar, speakers optimise their linguistic signals such that the surprisals $\iota_{\boldsymbol{w}}$ are distributed as uniformly as possible throughout a communication episode $\boldsymbol{w}$.

→ no evidence of local uniformity, pressure toward a global mean

→ information rate decreases in dialogues

Giulianelli & Fernández. CoNLL 2021.
Giulianelli, Sinclair, Fernández. AACL 2021.

Hypothesis 2: Structured Context

Values $\iota\left(w_t; \boldsymbol{w}_{<t}\right)$ in the surprisal contour $\iota_{\boldsymbol{w}}$ of a communication episode $\boldsymbol{w}$ are (partially) determined by the position of w_t within the hierarchy of $\boldsymbol{w}$'s constituent structural units.

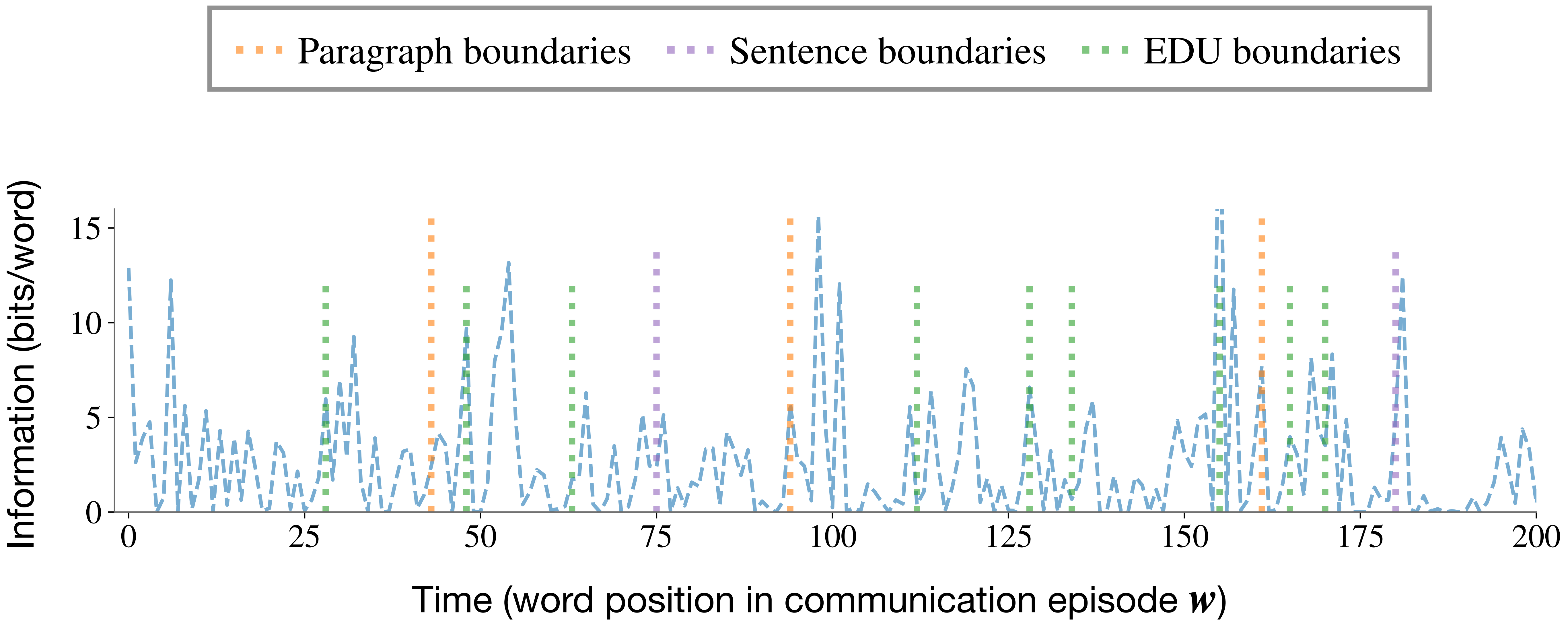
→ a unit's position within **contextual structure** predicts its surprisal

→ RST discourse units in texts

→ task-specific contextual units in dialogues

Giulianelli, Sinclair, Fernández. EMNLP 2021.
Tsipidi, Nowak, Cotterell, Wilcox, Giulianelli, Warstadt. EMNLP 2024.

Information contours in texts and dialogues



Hypothesis 1: Uniform Information Density

Subject to the constraints of the grammar, speakers optimise their linguistic signals such that the surprisals $\iota_{\boldsymbol{w}}$ are distributed as uniformly as possible throughout a communication episode $\boldsymbol{w}$.

→ no evidence of local uniformity, pressure toward a global mean

→ information rate decreases in dialogues

Giulianelli & Fernández. CoNLL 2021.
Giulianelli, Sinclair, Fernández. AACL 2021.

Hypothesis 2: Structured Context

Values $\iota\left(w_t; \boldsymbol{w}_{<t}\right)$ in the surprisal contour $\iota_{\boldsymbol{w}}$ of a communication episode $\boldsymbol{w}$ are (partially) determined by the position of w_t within the hierarchy of $\boldsymbol{w}$'s constituent structural units.

→ a unit's position within **contextual structure** predicts its surprisal

→ RST discourse units in texts

→ task-specific contextual units in dialogues

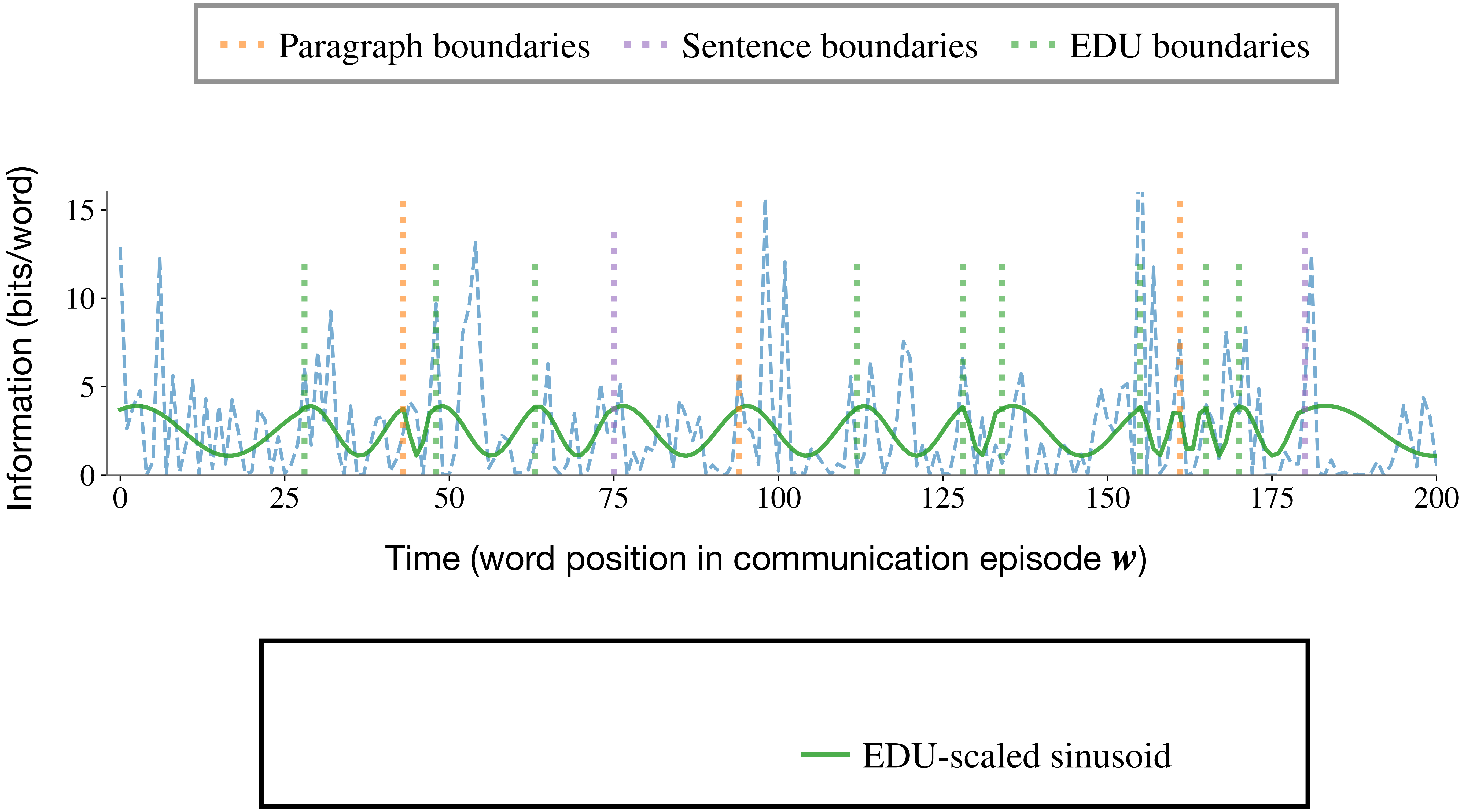
Giulianelli, Sinclair, Fernández. EMNLP 2021.
Tsipidi, Nowak, Cotterell, Wilcox, Giulianelli, Warstadt. EMNLP 2024.

Hypothesis 3: Harmonic Surprisal

Values $\iota\left(w_t; \boldsymbol{w}_{<t}\right)$ in the surprisal contour $\iota_{\boldsymbol{w}}$ of a communication episode $\boldsymbol{w}$ vary periodically, with periods that correspond to the boundaries of structural units within ι_{W} .

Tsipidi, Kiegeland, Nowak, Xu, Wilcox, Warstadt, Cotterell, Giulianelli. ACL 2025.

Information contours in texts and dialogues



Hypothesis 1: Uniform Information Density

Subject to the constraints of the grammar, speakers optimise their linguistic signals such that the surprisals $\iota_{\boldsymbol{w}}$ are distributed as uniformly as possible throughout a communication episode $\boldsymbol{w}$.

→ no evidence of local uniformity, pressure toward a global mean

→ information rate decreases in dialogues

Giulianelli & Fernández. CoNLL 2021.
Giulianelli, Sinclair, Fernández. AACL 2021.

Hypothesis 2: Structured Context

Values $\iota\left(w_t; \boldsymbol{w}_{<t}\right)$ in the surprisal contour $\iota_{\boldsymbol{w}}$ of a communication episode $\boldsymbol{w}$ are (partially) determined by the position of w_t within the hierarchy of $\boldsymbol{w}$'s constituent structural units.

→ a unit's position within **contextual structure** predicts its surprisal

→ RST discourse units in texts

→ task-specific contextual units in dialogues

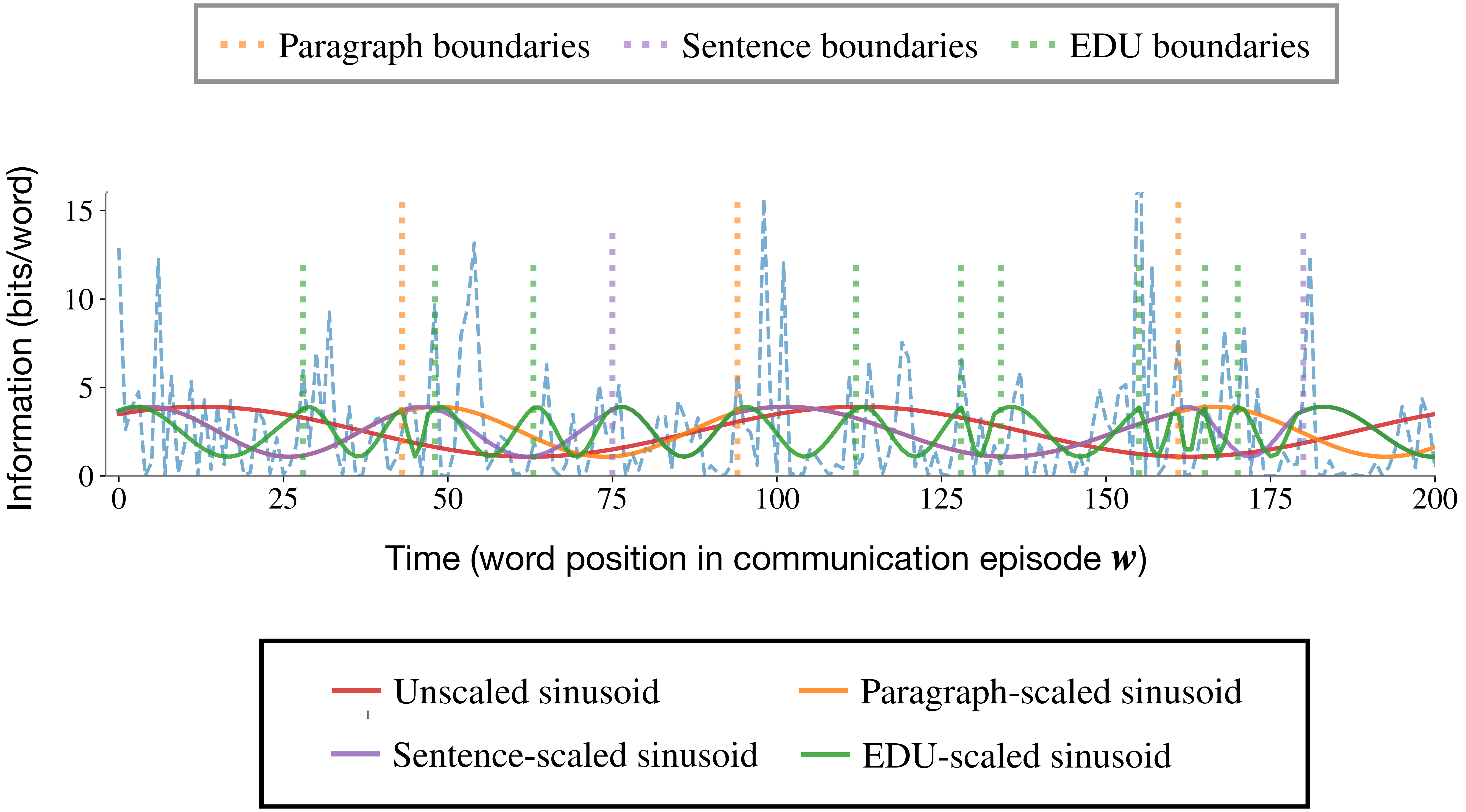
Giulianelli, Sinclair, Fernández. EMNLP 2021.
Tsipidi, Nowak, Cotterell, Wilcox, Giulianelli, Warstadt. EMNLP 2024.

Hypothesis 3: Harmonic Surprisal

Values $\iota\left(w_t; \boldsymbol{w}_{<t}\right)$ in the surprisal contour $\iota_{\boldsymbol{w}}$ of a communication episode $\boldsymbol{w}$ vary periodically, with periods that correspond to the boundaries of structural units within $\iota_{\boldsymbol{w}}$.

Tsipidi, Kiegeland, Nowak, Xu, Wilcox, Warstadt, Cotterell, Giulianelli. ACL 2025.

Information contours in texts and dialogues



Hypothesis 1: Uniform Information Density

Subject to the constraints of the grammar, speakers optimise their linguistic signals such that the surprisals $\iota_{\boldsymbol{w}}$ are distributed as uniformly as possible throughout a communication episode $\boldsymbol{w}$.

→ no evidence of local uniformity, pressure toward a global mean

→ information rate decreases in dialogues

Giulianelli & Fernández. CoNLL 2021.
Giulianelli, Sinclair, Fernández. AACL 2021.

Hypothesis 2: Structured Context

Values $\iota\left(w_t; \boldsymbol{w}_{<t}\right)$ in the surprisal contour $\iota_{\boldsymbol{w}}$ of a communication episode $\boldsymbol{w}$ are (partially) determined by the position of w_t within the hierarchy of $\boldsymbol{w}$'s constituent structural units.

→ a unit's position within **contextual structure** predicts its surprisal

→ RST discourse units in texts

→ task-specific contextual units in dialogues

Giulianelli, Sinclair, Fernández. EMNLP 2021.
Tsipidi, Nowak, Cotterell, Wilcox, Giulianelli, Warstadt. EMNLP 2024.

Hypothesis 3: Harmonic Surprisal

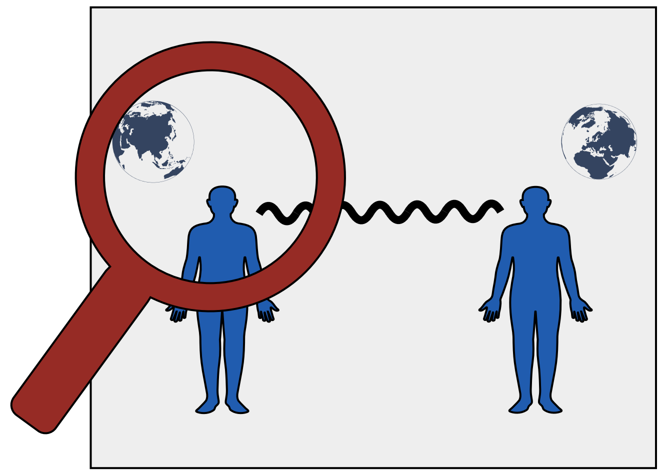
Values $\iota\left(w_t; \boldsymbol{w}_{<t}\right)$ in the surprisal contour $\iota_{\boldsymbol{w}}$ of a communication episode $\boldsymbol{w}$ vary periodically, with periods that correspond to the boundaries of structural units within $\iota_{\boldsymbol{w}}$.

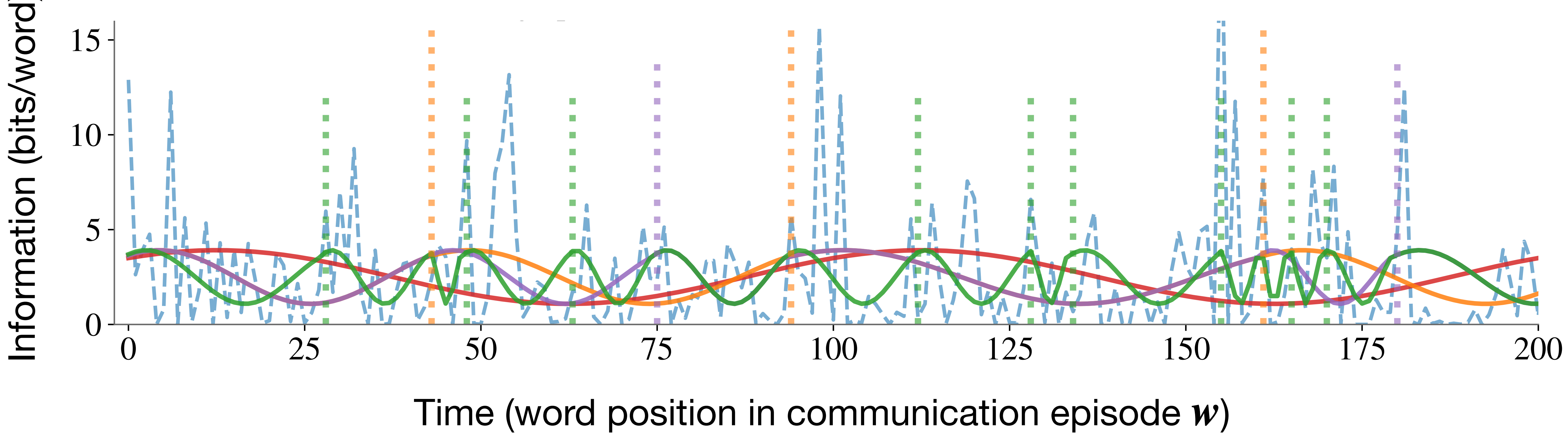
Tsipidi, Kiegeland, Nowak, Xu, Wilcox, Warstadt, Cotterell, Giulianelli. ACL 2025.

Information contours in texts and dialogues

Producers' communicative strategies through the lens of information rate modulation

- facilitating production (e.g., repetitions)
  Giulianelli, Sinclair, Fernández. AACL 2022.
- enhancing coordination in dialogue
  Yee, Giulianelli, Sinclair. LREC-COLING 2024.
- style, genre, and writing quality
- human-generated vs. model-generated texts
- facilitating comprehension
- in multimodal contexts
  Gay, Haley, Giulianelli, Ponto. EACL 2026.





Hypothesis 1: Uniform Information Density

Subject to the constraints of the grammar, speakers optimise their linguistic signals such that the surprisals $\iota_{\boldsymbol{w}}$ are distributed as uniformly as possible throughout a communication episode $\boldsymbol{w}$.

→ no evidence of local uniformity, pressure toward a global mean
→ information rate decreases in dialogues

Giulianelli & Fernández. CoNLL 2021.
Giulianelli, Sinclair, Fernández. AACL 2021.

Hypothesis 2: Structured Context

Values $\iota\left(w_t; \boldsymbol{w}_{<t}\right)$ in the surprisal contour $\iota_{\boldsymbol{w}}$ of a communication episode $\boldsymbol{w}$ are (partially) determined by the position of w_t within the hierarchy of $\boldsymbol{w}$'s constituent structural units.

→ a unit's position within **contextual structure** predicts its surprisal
  → RST discourse units in texts
  → task-specific contextual units in dialogues

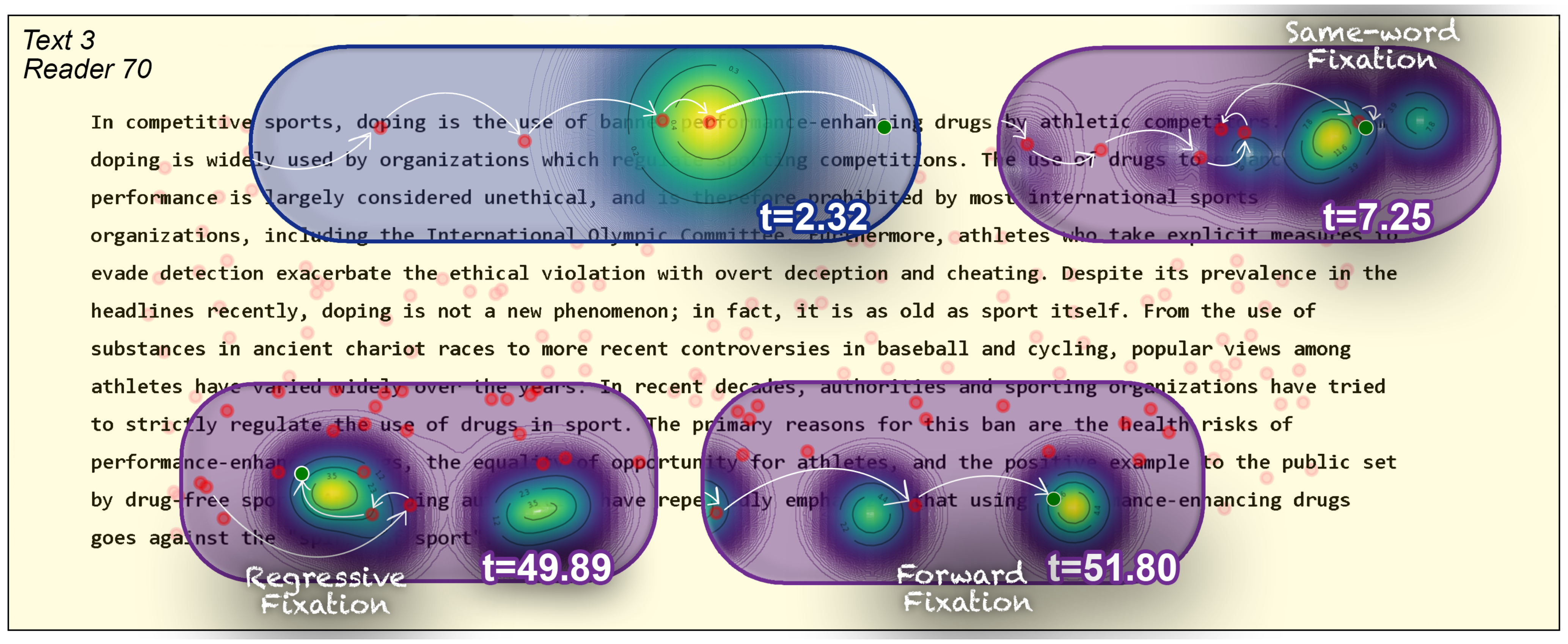
Giulianelli, Sinclair, Fernández. EMNLP 2021.
Tsipidi, Nowak, Cotterell, Wilcox, Giulianelli, Warstadt. EMNLP 2024.

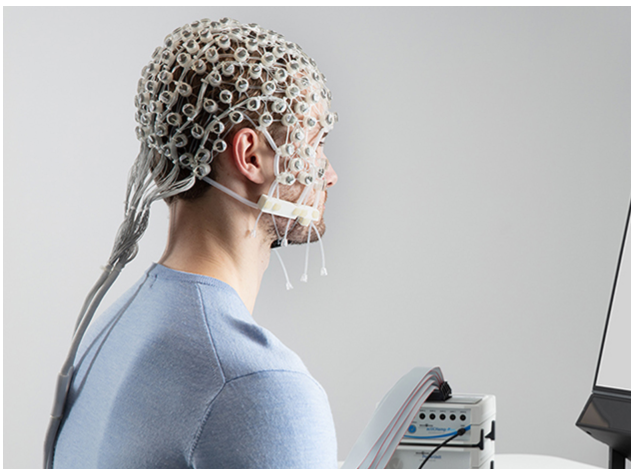
Hypothesis 3: Harmonic Surprisal

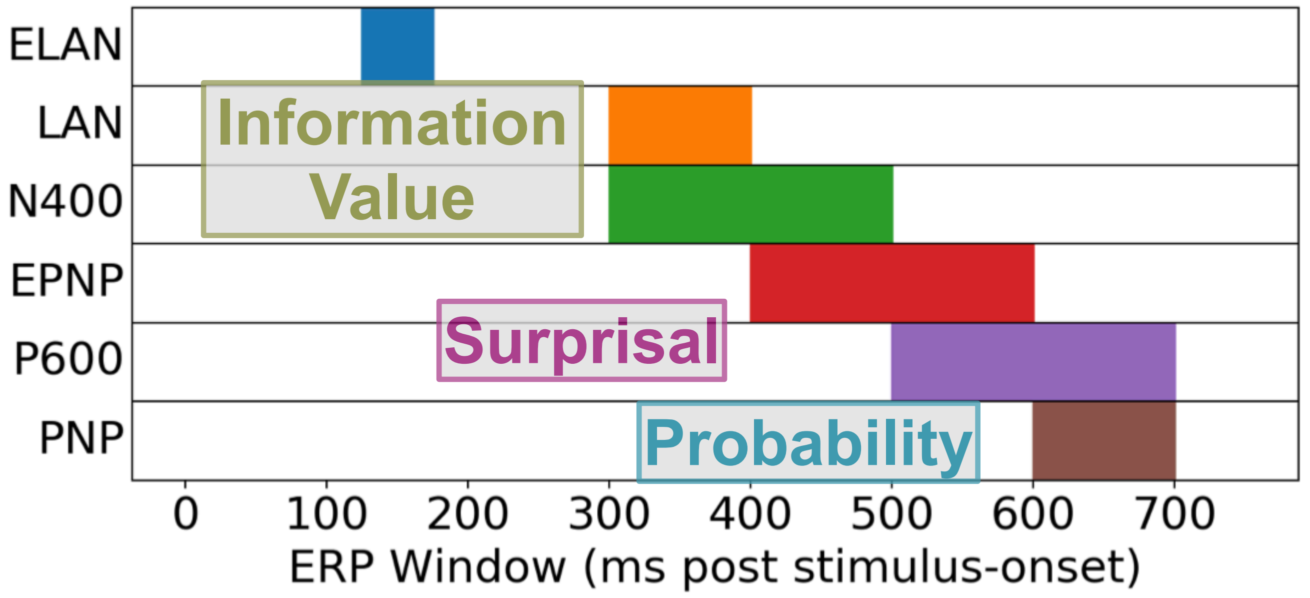
Values $\iota\left(w_t; \boldsymbol{w}_{<t}\right)$ in the surprisal contour $\iota_{\boldsymbol{w}}$ of a communication episode $\boldsymbol{w}$ vary periodically, with periods that correspond to the boundaries of structural units within $\iota_{\boldsymbol{w}}$.

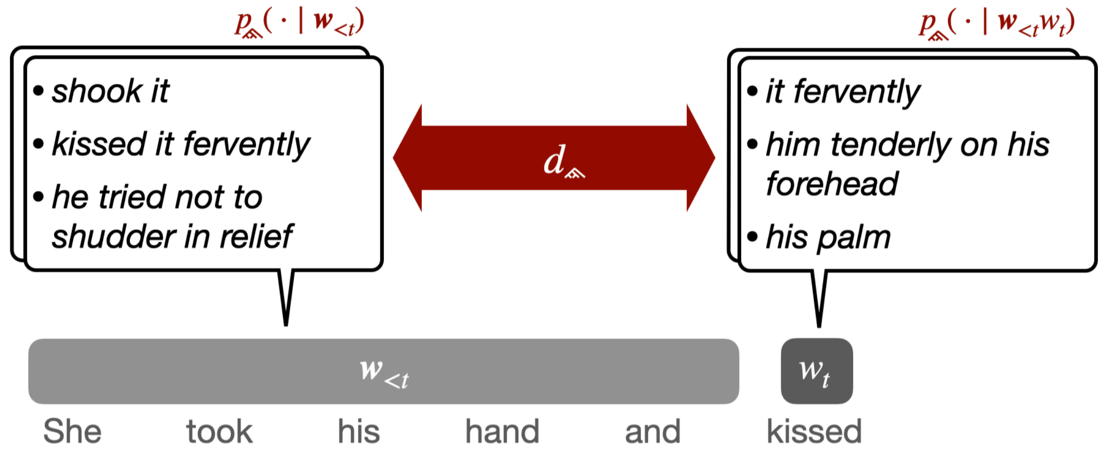
Tsipidi, Kiegeland, Nowak, Xu, Wilcox, Warstadt, Cotterell, Giulianelli. ACL 2025.

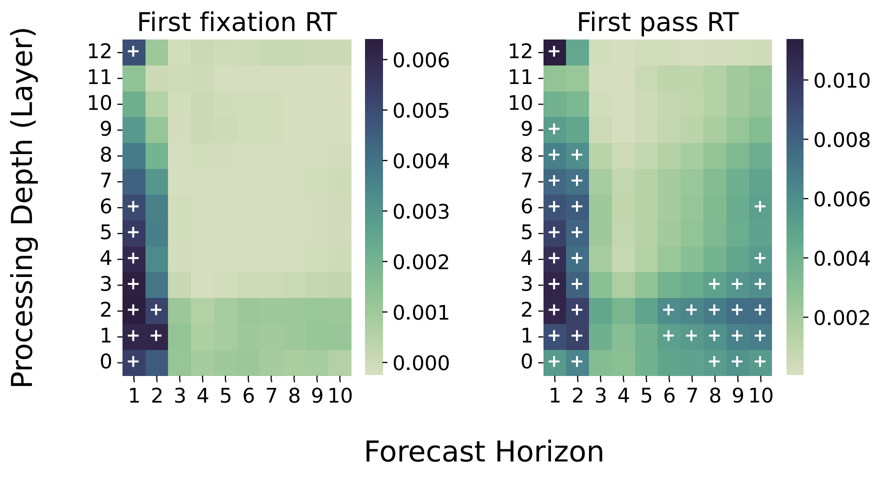
Comprehension



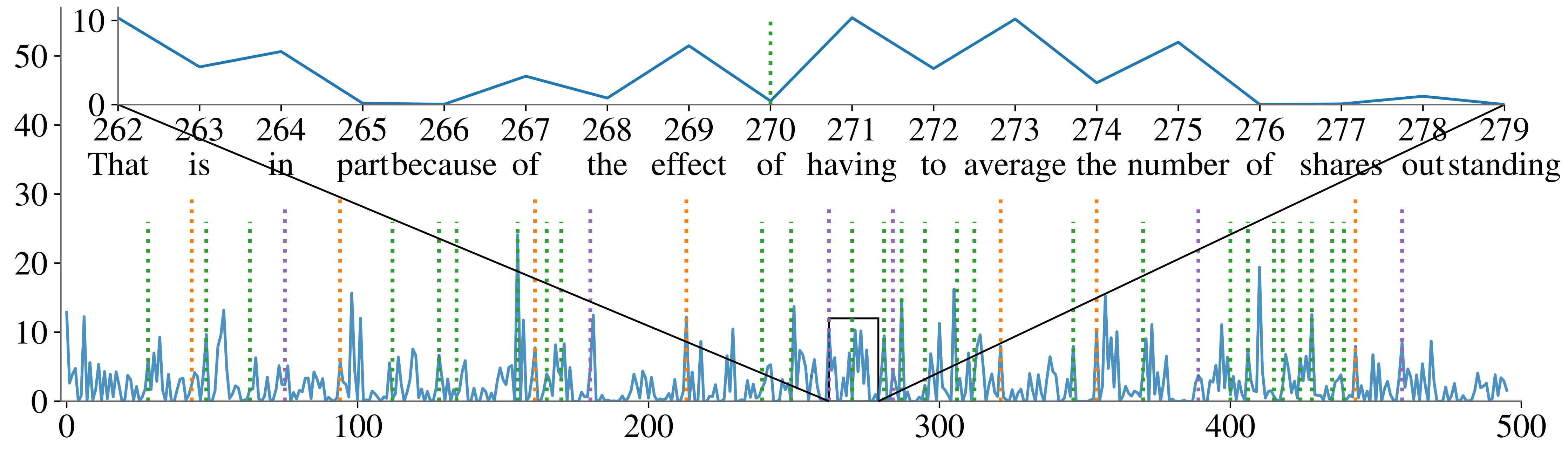


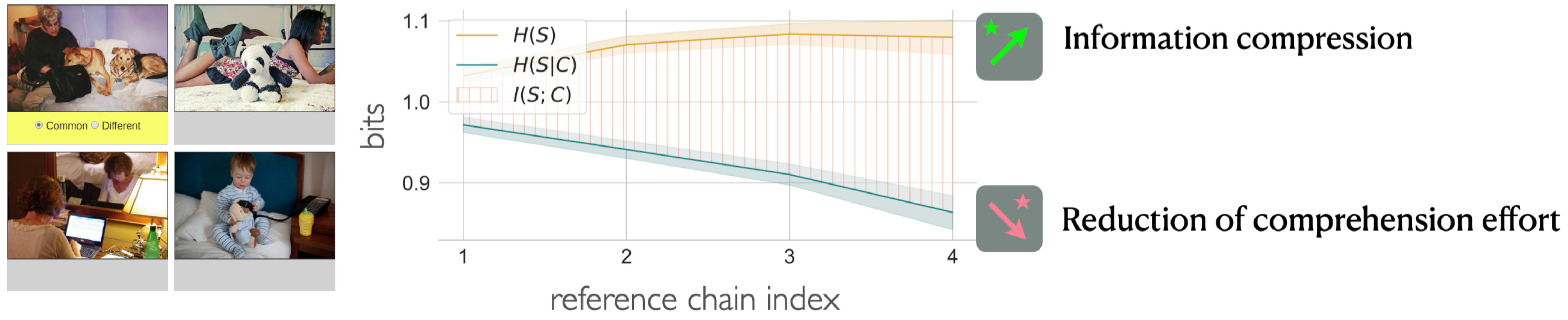


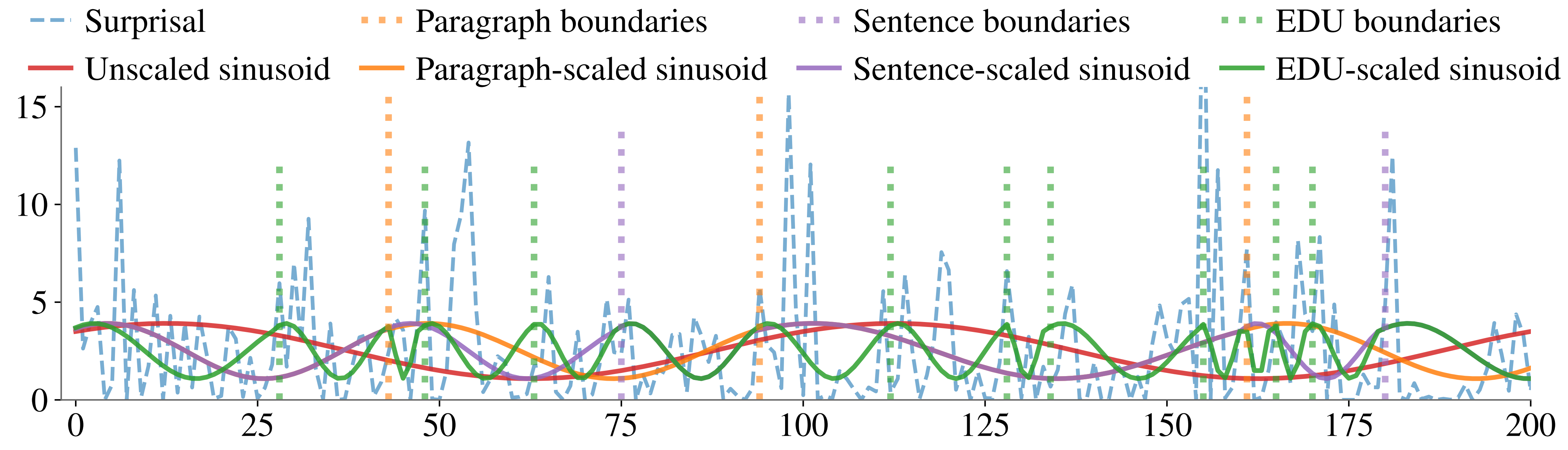




Production

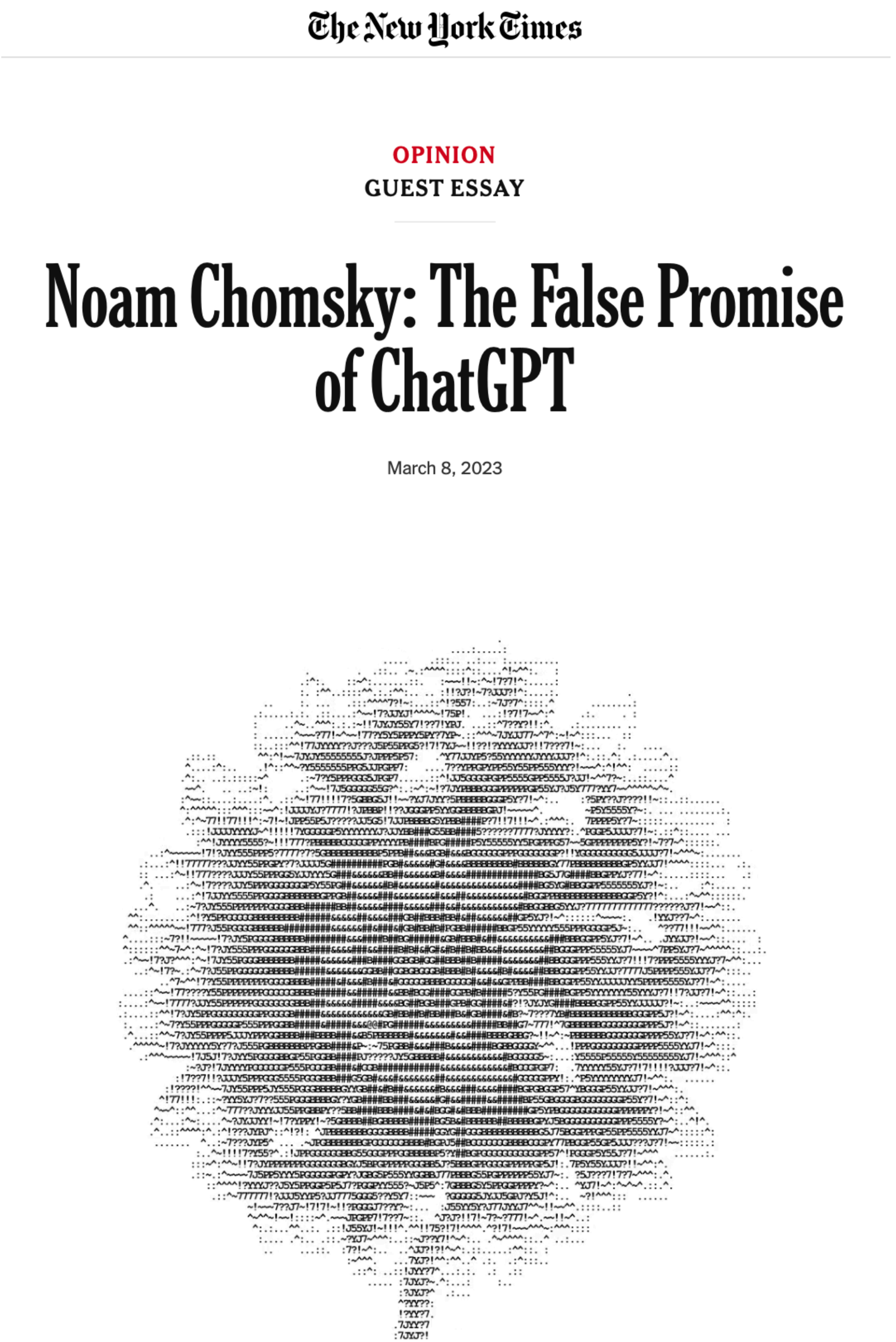






Information-theoretic predictors of language learnability

Do neural networks and humans share similar inductive biases?

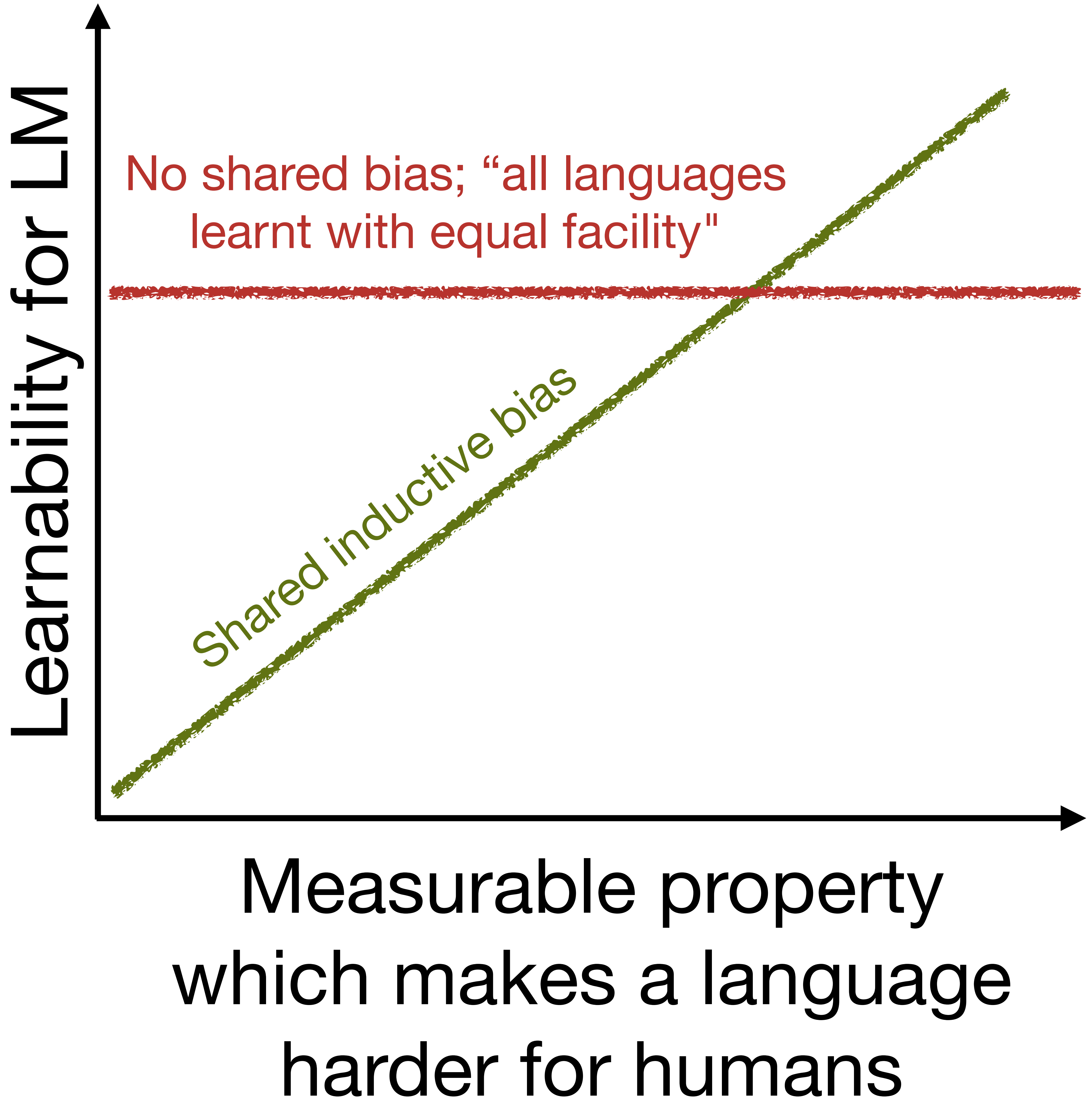


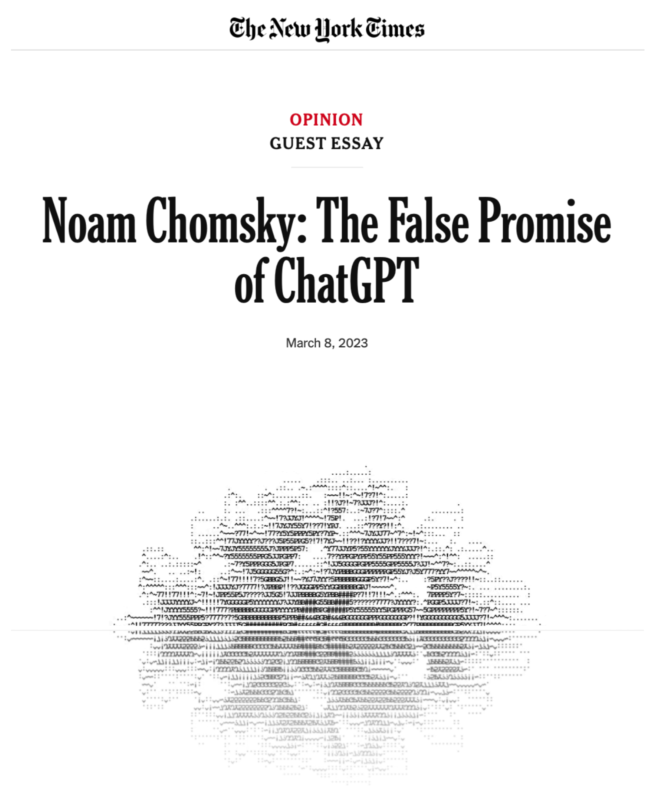
The New York Times

OPINION
GUEST ESSAY

Noam Chomsky: The False Promise of ChatGPT

March 8, 2023




The New York Times

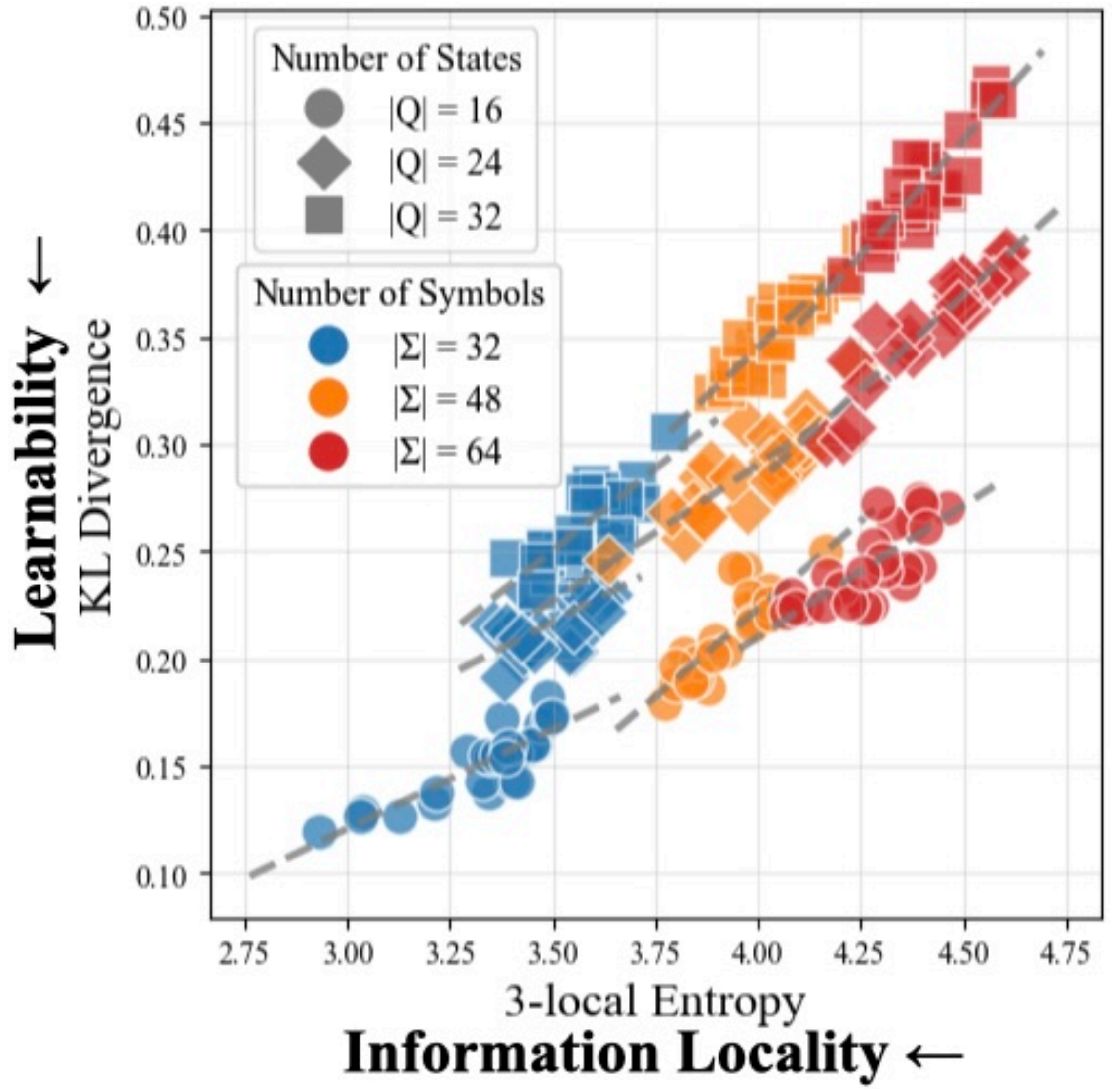
OPINION
GUEST ESSAY

Noam Chomsky: The False Promise of ChatGPT

March 8, 2023

Information-theoretic predictors of language learnability

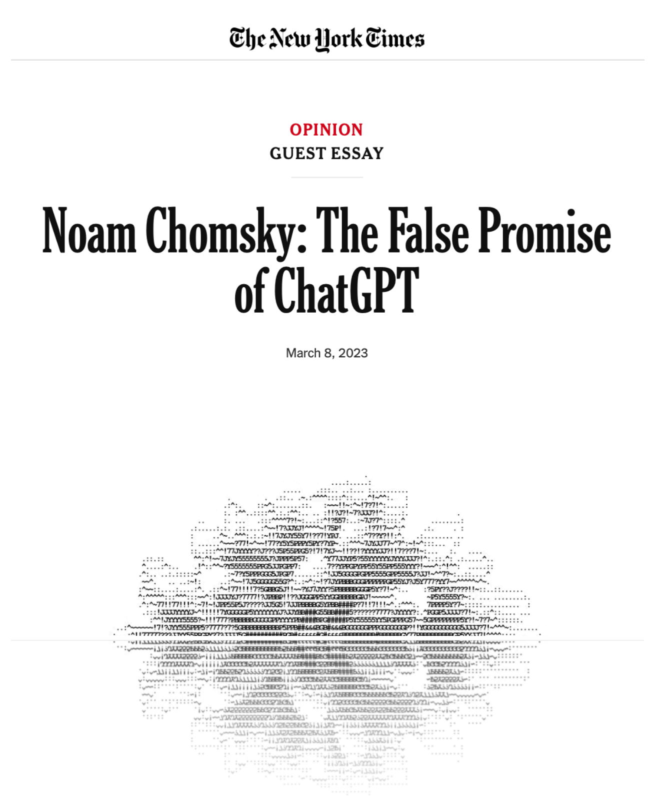
Do neural networks and humans share similar inductive biases?



<u>First empirical findings</u>

Learnability in LSTMs and Transformers is predicted by **information locality** (Gibson 2001; Futrell et al. 2020)

Someya, Svete, DuSell, O'Donnell, Giulianelli, Cotterell. ACL 2025.


The New York Times

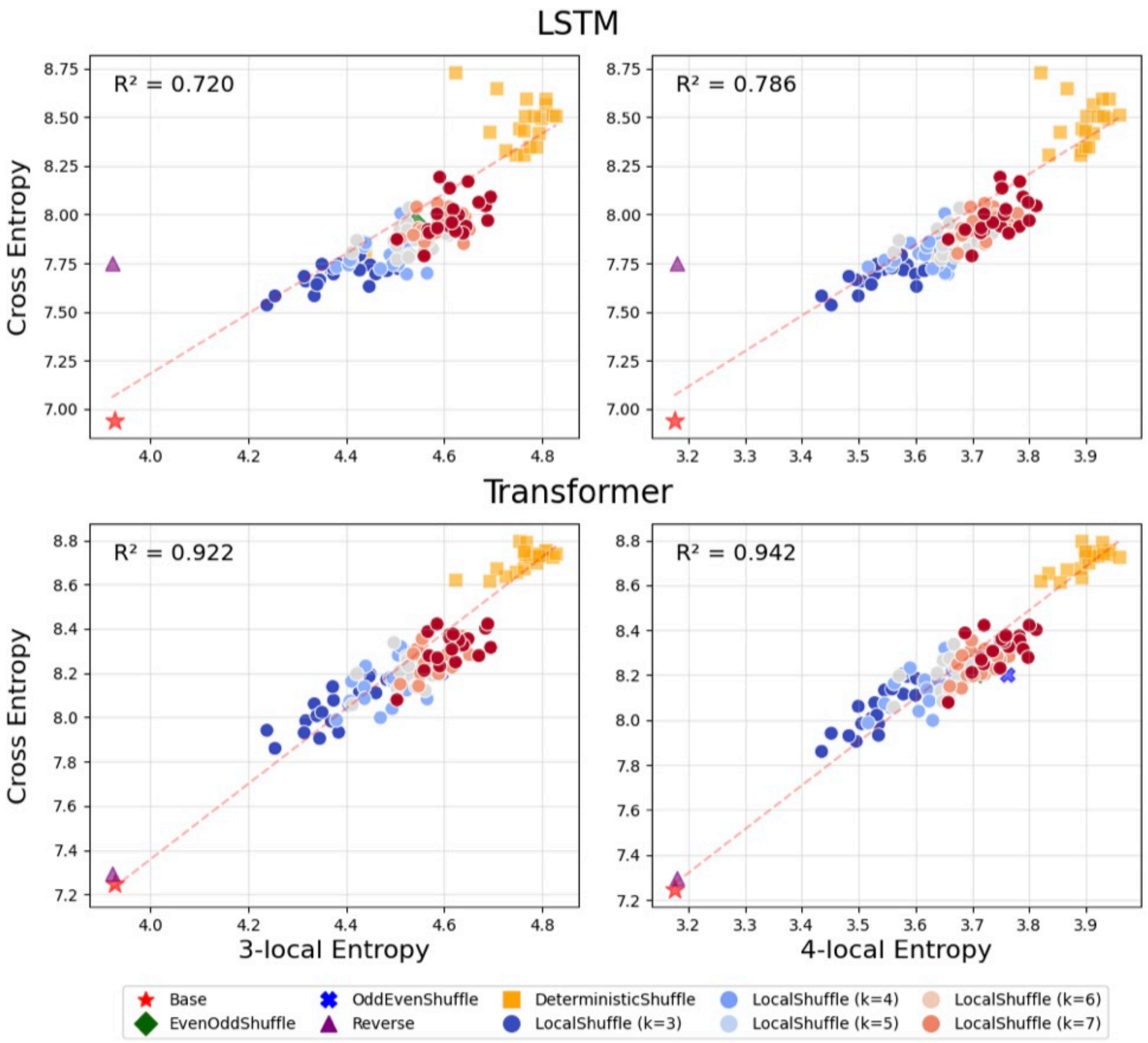
OPINION
GUEST ESSAY

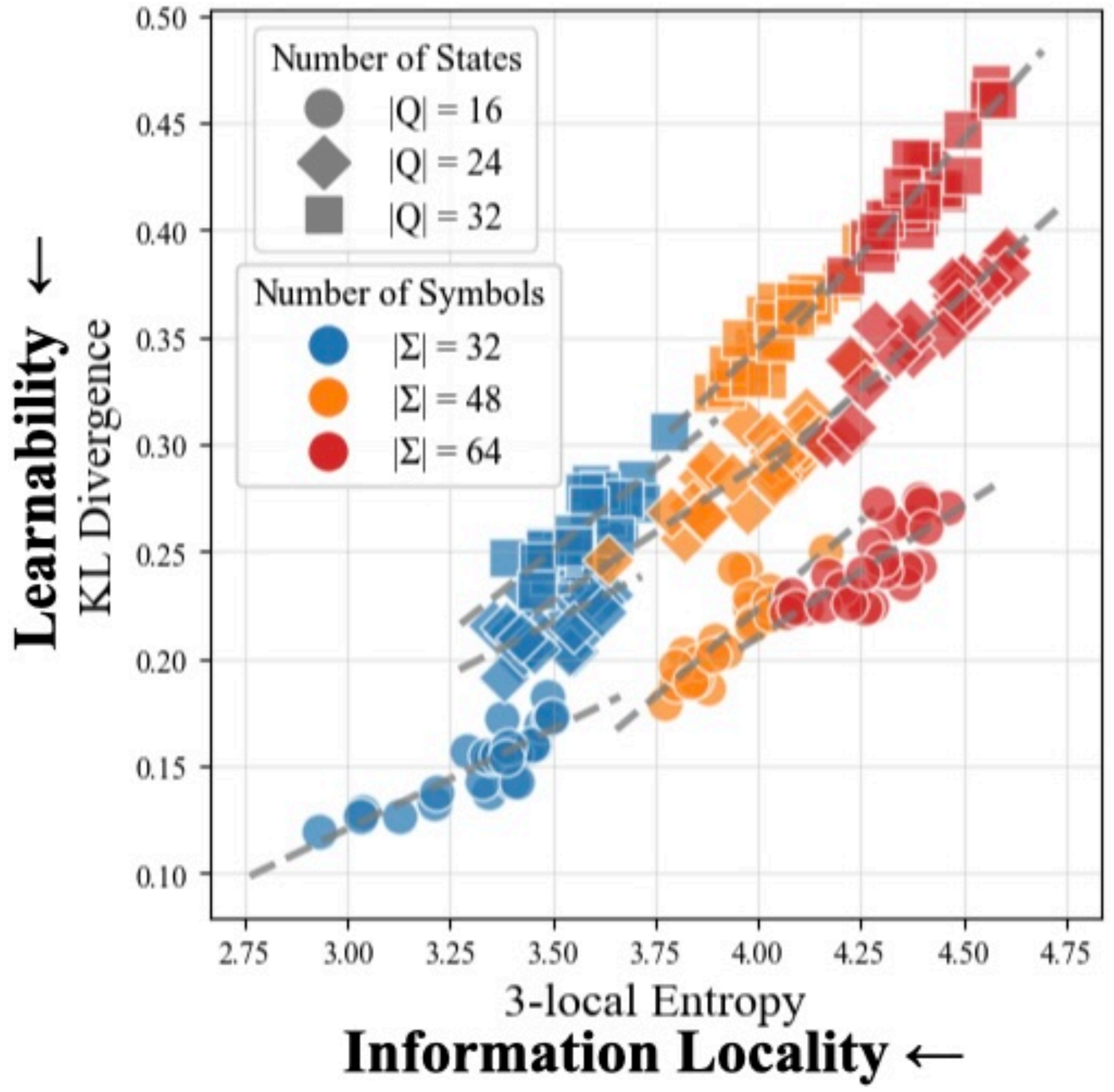
Noam Chomsky: The False Promise of ChatGPT

March 8, 2023

Information-theoretic predictors of language learnability

Do neural networks and humans share similar inductive biases?





<u>First empirical findings</u>

Learnability in LSTMs and Transformers is predicted by **information locality** (Gibson 2001; Futrell et al. 2020)

Someya, Svete, DuSell, O'Donnell, Giulianelli, Cotterell. ACL 2025.

A case for optimism

LMs as tools to run **computational simulations** of language processing in humans**.**

Information theory as an interlingua to express theoretical constructs and formulate *executable hypotheses*.

LMs **enable and accelerate the refinement of scientific hypotheses** concerning language comprehension, production, and learning.

Beyond Language

Richer contextualisation, multi-modality, and the ability to act and interact in *AI agents* open new avenues for psychology, neuroscience, and the broader social sciences.