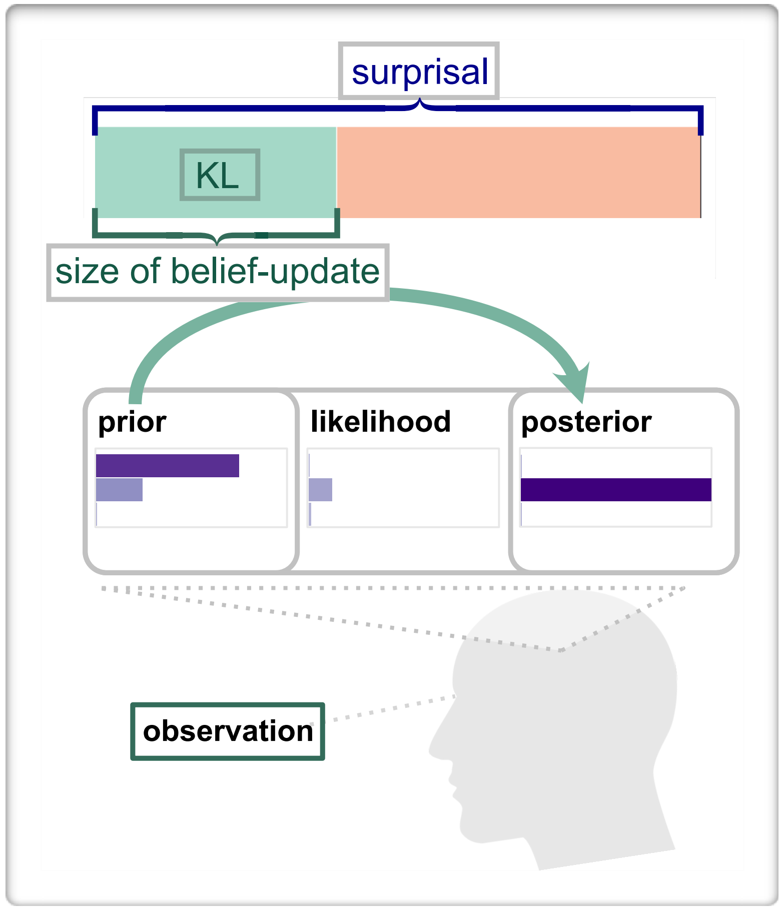
incremental processing difficulty as the cost of inference

Jacob Hoover Vigly
MIT

7 May 2025, Language Processing Group at UCI



slides: jahoo.github.io/2025-05-07

language comprehension

how do we understand what a sentence means?

meaning $z =$ 

inference

performed under uncertainty

utterance $\mathbf{u} =$ Near that bank there is an otter

- perceive utterance, $\mathbf{u} = u_1, u_2, \ldots$ word by word
- infer meaning, z, in context

language comprehension

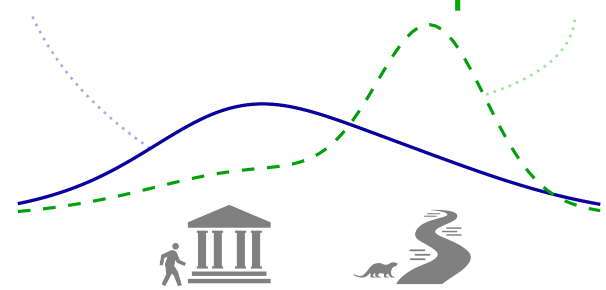
iterative inference problem

$z =$ 

$\mathbf{u}$ = Near that bank htere is an otter …

- observe otterance word by word $\mathbf{u} = u_1, u_2, \ldots$ in noisy environment
- with each word, **change beliefs** about meaning, z

u_i causes belief update $\underbrace{p(Z \mid u_{1\ldots i-1})}_{\text{prior}} \overset{u_i}{\rightsquigarrow} \underbrace{p(Z \mid u_{1\ldots i-1}, u_i)}_{\text{posterior}}$

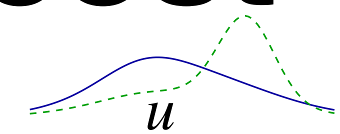


How? …with what **processing algorithm**?

- **important clue** for humans, **unexpected words take more effort**.
- **intuition** bigger update = higher cost

incremental processing cost

How? …with what processing algorithm?



$p_Z \overset{u}{\mapsto} p_{Z|u}$

- important clue: for humans, **unexpected words take more effort**.

has been formalized as

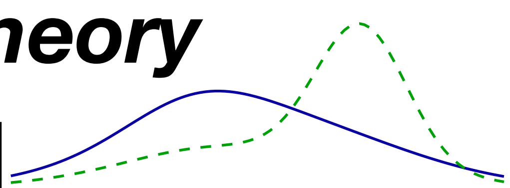
surprisal theory
(Hale '01, Levy '08)

$$\mathrm{cost}(u) \propto \overbrace{\log \frac{1}{p(u)}}^{\mathrm{surprisal}(u)}$$

precise formalization of hypothesis… but how? what algorithm?

- idea: difficult = big update (resource allocation cost)

update-size *theory*

$\mathrm{cost}(u) = f(\|$  $\|)$

size of belief update

hypothesis that cost measured as **<u>bits of information gained</u>** about Z

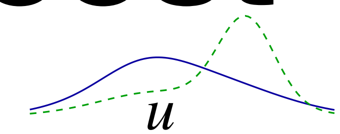
surprisal theory is special case, by two assumptions:

(a) that $\mathrm{D}(p_{Z|u} \| p_Z) = \mathrm{surprisal}$ (extra term is zero) ⟵ will focus on this later

(b) that f is linear

incremental processing cost

How? …with what processing algorithm?



$p_Z \overset{u}{\mapsto} p_{Z|u}$

- important clue: for humans, **unexpected words take more effort**.
- intuition: bigger update = higher cost

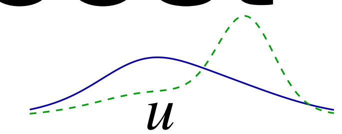
many candidate algorithms don't have this property

- parsing algorithms (Z ranges over trees)
  - non-probabilistic algorithms
  - probabilistic enumerative algorithms
  - neural-parametrized parsing algorithms
- language model inference (e.g. n-gram, RNN, Transformer LLMs …)

… amount of work done during inference **doesn't depend on probabilistic properties at all**

(so, they don't directly explain this human behavior)

incremental processing cost



How? …with what processing algorithm? $p_Z \overset{u}{\mapsto} p_{Z|u}$

- important clue: for humans, **unexpected words take more effort**.
- intuition: bigger update = higher cost

what kind of algorithms *do* have this property?

those that somehow prioritize more probable hypotheses:

- sampling algorithms

➡ e.g. *rejection sampling* - guess-and-check until success

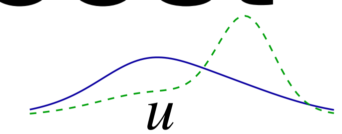
$$\mathbb{E}\ \#\text{samples} = 1 \Big/ \Pr(\text{success})$$

$$= 1 \Big/ \textstyle\sum_z p(z)\underbrace{p(u \mid z)}_{\text{lik}(z)} = 1 \Big/ p(u)$$

$$= e^{-\log p(u)} = e^{\text{surprisal}(u)}$$

incremental processing cost

How? …with what processing algorithm?



$$p_Z \overset{u}{\mapsto} p_{Z|u}$$

- important clue: for humans, **unexpected words take more effort**.
- intuition: bigger update = higher cost

what kind of algorithms *do* have this property?

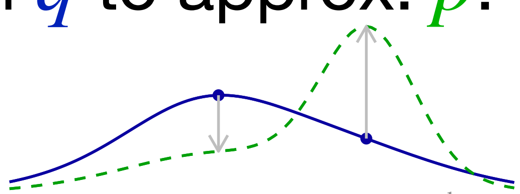
those that somehow prioritize more probable hypotheses:

- sampling algorithms

➡ *importance sampling* complexity scales in **divergence**:

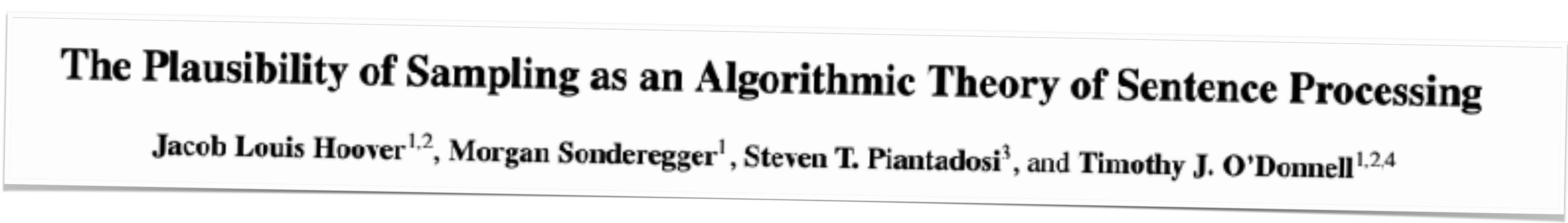
sampling from q to approx. p: req #samples $\approx e^{\mathrm{D}_{\mathrm{KL}}(p\|q)}$ Chatterjee & Diaconis 2018, …

$\approx \mathrm{D}_{\chi^2}(p\|q)$ Agapiou et al. 2017, Sanz-Alonso 2018, …



importance weight $w(z) \propto \frac{\mathrm{d}p}{\mathrm{d}q}(z)$

$$\mathrm{cost}(u) = f\big(\ \mathrm{D}(p_{Z|u}\|p_Z)\ \big)$$



The Plausibility of Sampling as an Algorithmic Theory of Sentence Processing

Jacob Louis Hoover[1,2], Morgan Sonderegger[1], Steven T. Piantadosi[3], and Timothy J. O'Donnell[1,2,4]

The Plausibility of Sampling as an Algorithmic Theory of Sentence Processing

Jacob Louis Hoover[1,2], Morgan Sonderegger[1], Steven T. Piantadosi[3], and Timothy J. O'Donnell[1,2,4]

What kinds of mechanisms prioritize high-probability hypotheses?

- **sampling** according to probability
  - simple rejection **sampling**
  - rejection **sampling** w/o replacement
  - importance **sampling**
- **ranked search** in order of probability

$$
\begin{aligned}
\mathrm{cost}(u) &= f\left(\ D(p_{Z|u} \| p_Z)\ \right)\\
&= f(\mathrm{surprisal}(u)) \quad \text{assumption (a)} \longleftarrow \text{for this paper, we assumed this}\\
&\propto \ \mathrm{surprisal}(u) \quad \text{assumption (b)} \longleftarrow \text{focused on this}
\end{aligned}
$$

sampling algorithms predict:

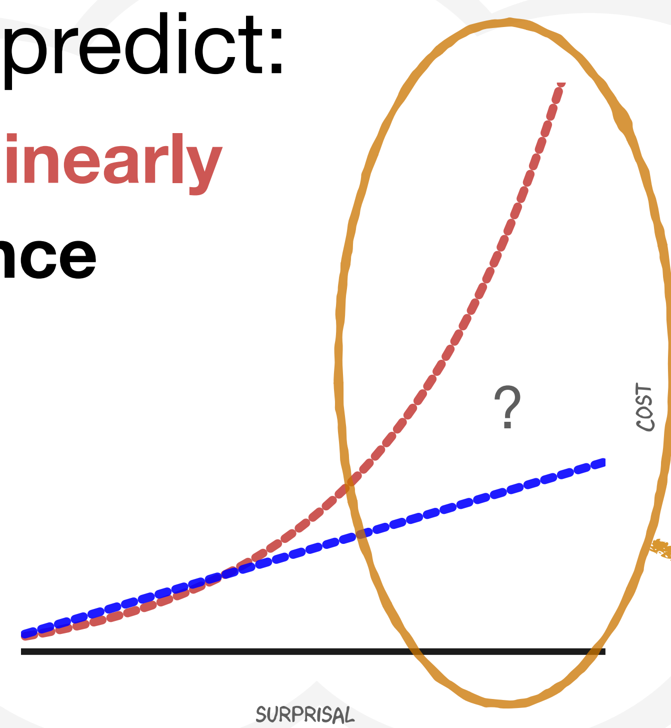
⇒ cost increases **super-linearly**

⇒ with **increasing variance**

but surprisal theory proposes

⇒ cost increases **linearly**

⇒ (says nothing about variance)



high-surprisal region particularly important

The Plausibility of Sampling as an Algorithmic Theory of Sentence Processing

Jacob Louis Hoover[1,2], Morgan Sonderegger[1], Steven T. Piantadosi[3], and Timothy J. O'Donnell[1,2,4]

linking function: empirical study

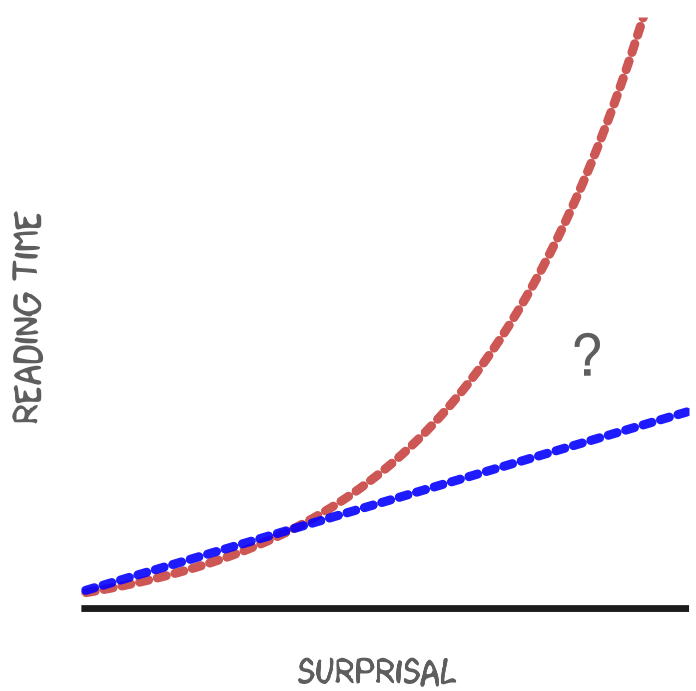
is the mean superlinear? does variance increase?

fit **location-scale Generalized Additive Model (GAMs)**

- potential **nonlinear effect of surprisal** on RT
- likewise on **variance** in RT

predictor of interest: surprisal

- estimate with **pretrained LLMs**



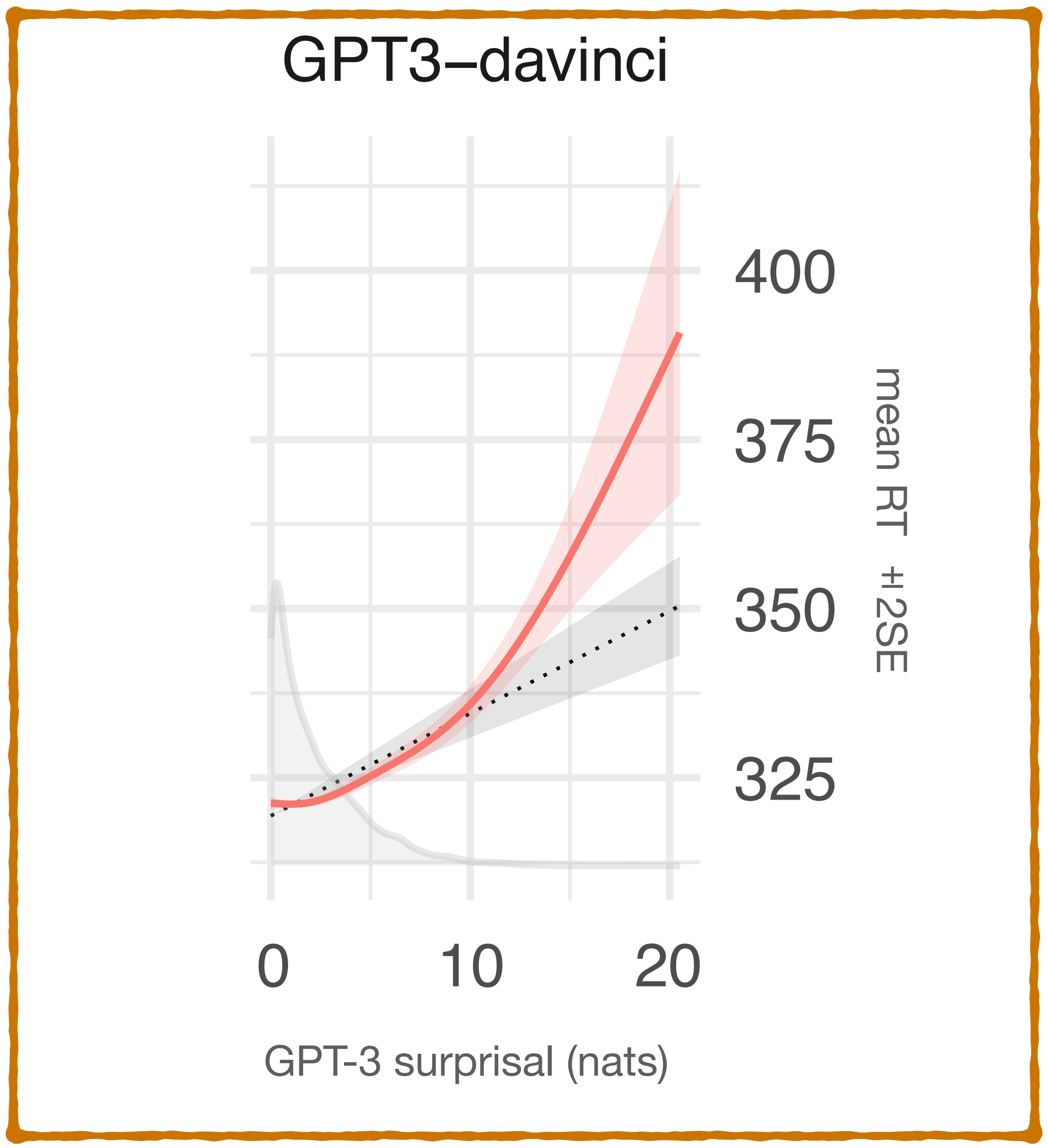
response: processing time

- self-paced reading time
- used Natural Stories data set
  - 10 stories, ~1000 words each
  - RTs from avg 84 participants
  - containing rare constructions (wide range of surprisals helpful to **distinguish linking function**)

linking function: empirical study

is the mean superlinear? does variance increase?

Yes



The Plausibility of Sampling as an Algorithmic Theory of Sentence Processing

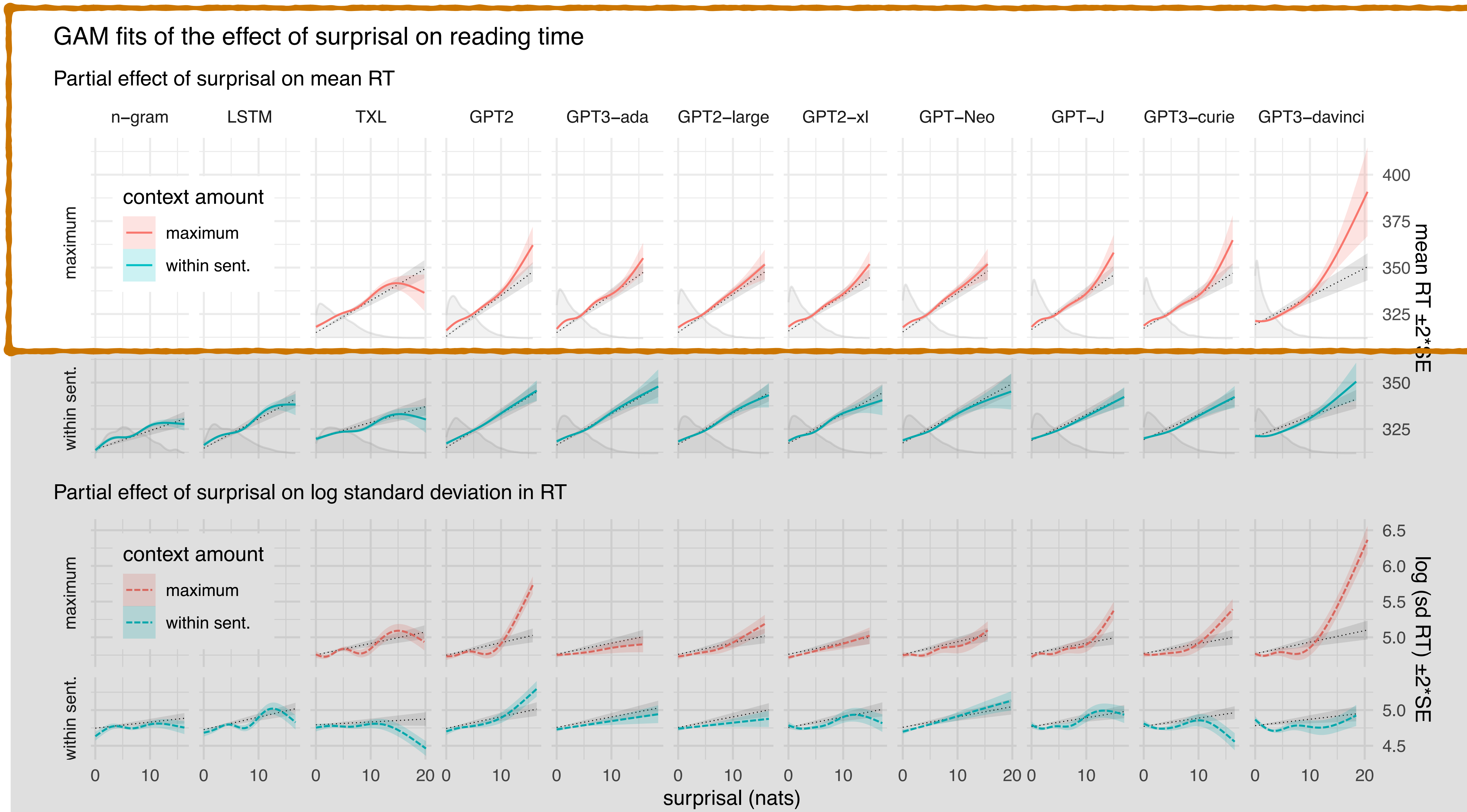
Jacob Louis Hoover[1,2], Morgan Sondereggerr[1], Steven T. Piantadosi[3], and Timothy J. O'Donnell[1,2,4]

linking function: empirical study

is the mean superlinear? does variance increase?

Yes

- better LM ⇒ more superlinear



The Plausibility of Sampling as an Algorithmic Theory of Sentence Processing

Jacob Louis Hoover[1,2], Morgan Sonderegger[1], Steven T. Piantadosi[3], and Timothy J. O'Donnell[1,2,4]

linking function: empirical study

is the mean superlinear?

Yes

- better LM $\Rightarrow$ more superlinear

linear surprisal theory
(Hale '01, Levy '08)

$$\mathrm{cost}(u) \propto \mathrm{surprisal}(u)$$

***general** surprisal theory*
(Levy '05, Meister '21, Xu '23)

$$\mathrm{cost}(u) = f(\mathrm{surprisal}(u))$$

does variance increase?

Yes

- across LMs

consistent with sampling algorithms' predictions

$\Longrightarrow$ **motivation:** sampling mechanisms for processing

when surprisal ≠ divergence

but, there was another assumption! …that surprisal = divergence

surprisal theory

$$\text{cost}(u) = f(\text{surprisal}(u))$$

belief-update theory

$$\boxed{\text{cost}(u) = f(\mathrm{D}_{\mathrm{KL}}(p_{Z|u} \| p_Z))}$$

recall motivation: surprisal as measure of size of belief update

$$\overbrace{\mathrm{D}_{\mathrm{KL}}(p_{Z|u} \| p_Z)}^{\mathbb{E}_{p_{Z|u}}\left[\log \frac{p(z \mid u)}{p(z)}\right]} = \overbrace{\text{surprisal}(u)}^{\log \frac{1}{p(u)}} - \overbrace{R(u)}^{\mathbb{E}_{p_{Z|u}}\left[\log \frac{1}{p(u \mid z)}\right]}$$

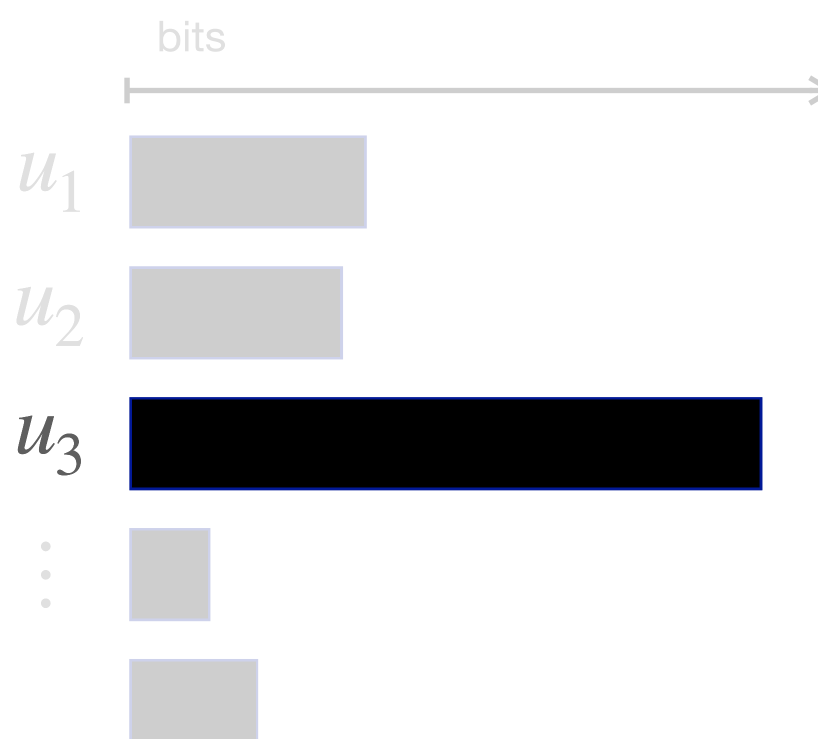
when surprisal > KL divergence

raw amount of info contained in u

size of belief update caused by observing u

reconstruction information: 'extra' bits that don't contribute to update

$$\text{surprisal} = D_{\mathrm{KL}} + R$$

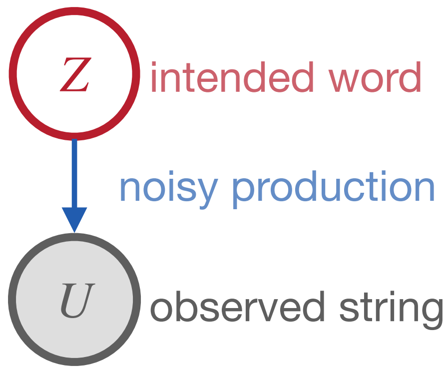


Comprehension difficulty as the cost of inference
(WIP with Peng Qian, Morgan Sonderegger, Tim O'Donnell)

typos as a case study
when unpredictable doesn't mean difficult

typos as a case study

surprisal = $\mathrm{D}_{\mathrm{KL}} + R$



(chosen from some set $\mathcal{Z}$ = `wordlist`)

For now,

let latent Z (meaning) range over strings, representing **intended word**

- easy to model prior and likelihood
- narrow application where we might expect **LM surprisal of the observed string is intuitively inadequate** as measure of human processing cost
- (Note: I'm interested in broader applications to follow!)

typos as a case study

surprisal $= D_{KL} + R$

Example:

- *After tripping on the rug and falling in front of everyone, I felt deeply* ______

condition	target word	surprisal	divergence	
1. expected	*embarrassed*	LOW	LOW	😐
2. unexpected	*innovative*	HIGH	HIGH	🤯
3. expected (typo)	*embarrsased*	HIGH	LOW	😕
4. unexpected (typo)	*innovaitve*	HIGH	HIGH	

(with *any* plausible noise model)

typos as a case study

$$\text{surprisal} = D_{KL} + R$$

Example:

- *After tripping on the rug and falling in front of everyone, I felt deeply* ______

1. expected — *embarrassed*
2. unexpected — *innovative*
3. expected (typo) — *embarrsased*
4. unexpected (typo) — *innovaitve*

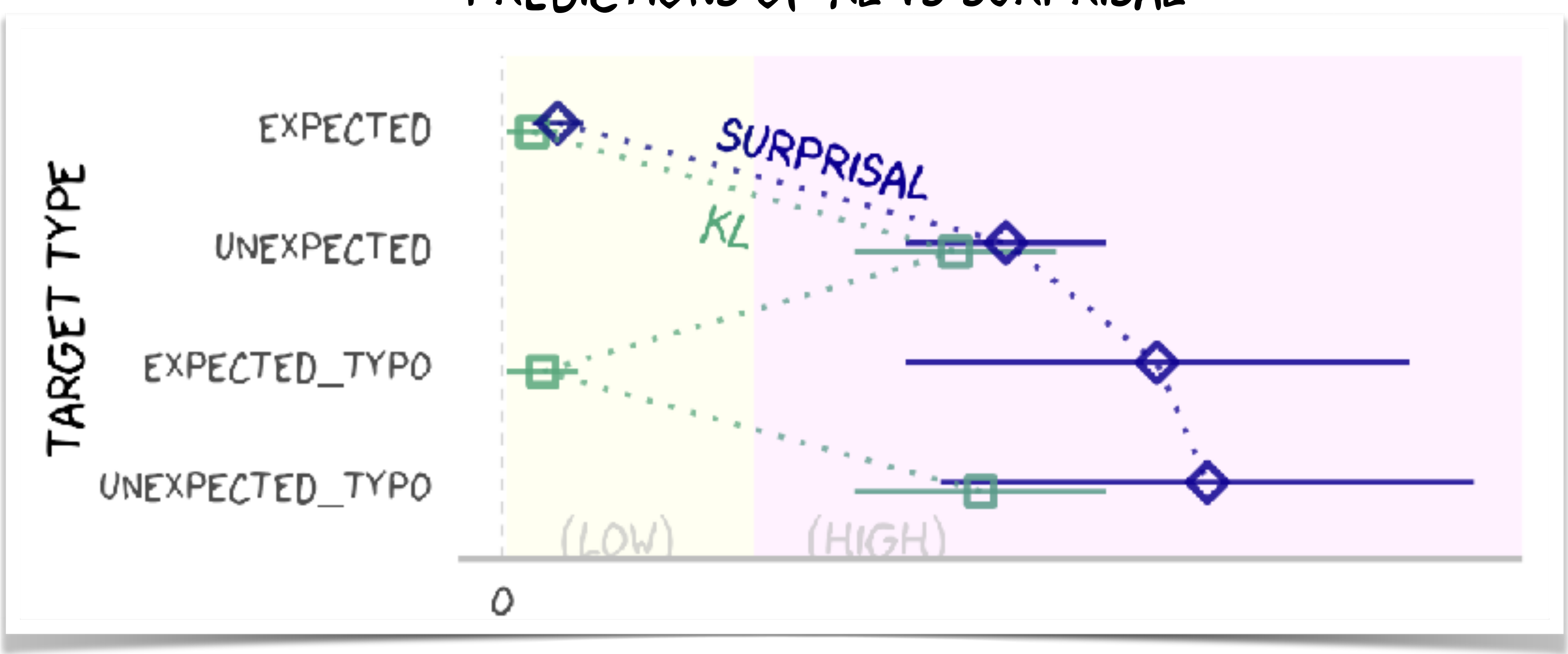
Self-paced reading time study:

- 51 sentences x 4 conditions = 204 unique targets of interest.
- 104 participants on Prolific (post exclusions)

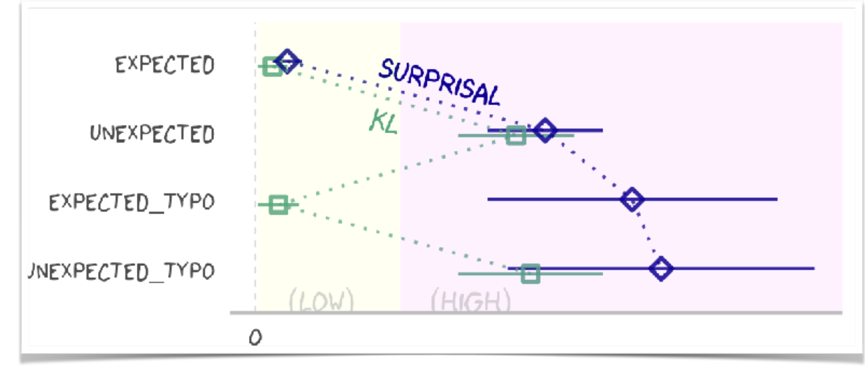
Fit mixed-effect regression models:

- predict human RT
- predict LLM surprisal (separately)
  - surprisals from collection of LLMs

PREDICTIONS OF KL VS SURPRISAL



typos as a case study



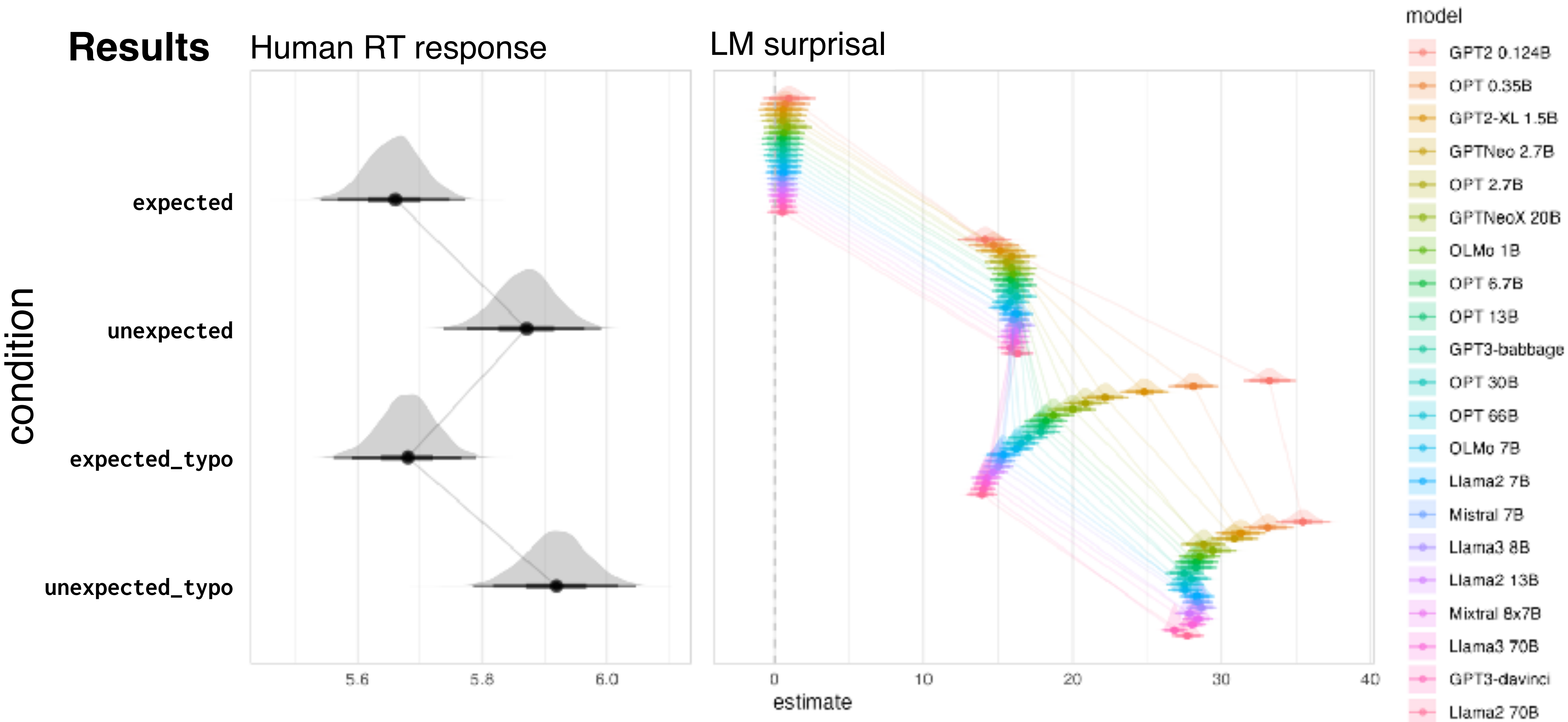
Does surprisal pattern as expected?

Yes. Surprisal is low in **`expected`** condition, but high in others.

Does human RT pattern like surprisal or divergence?

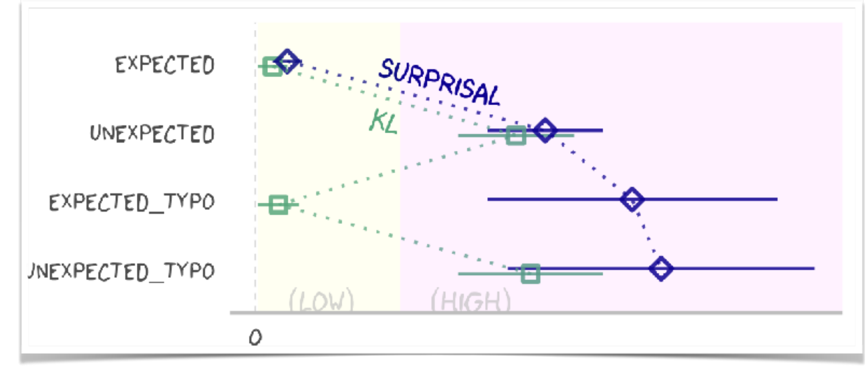
RTs zig-zag, **as divergence should predict**, contra surprisal.

Results

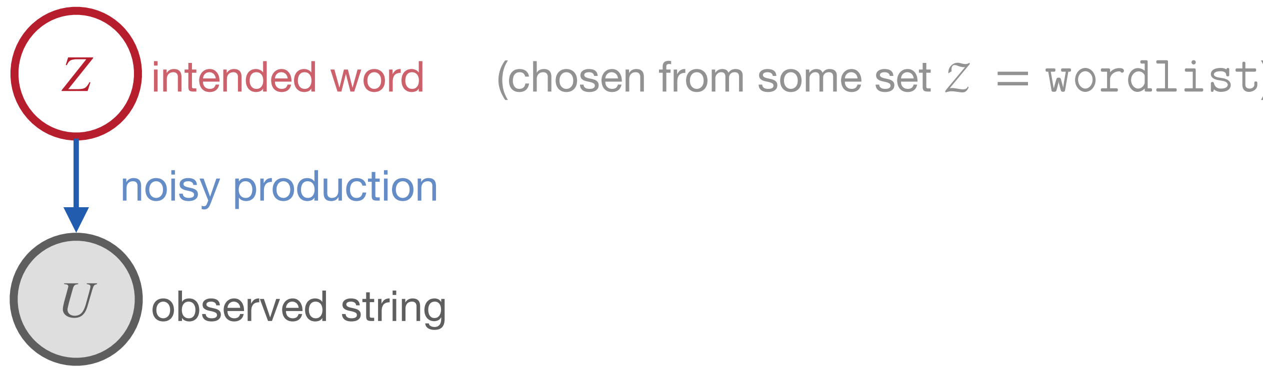


typos as a case study

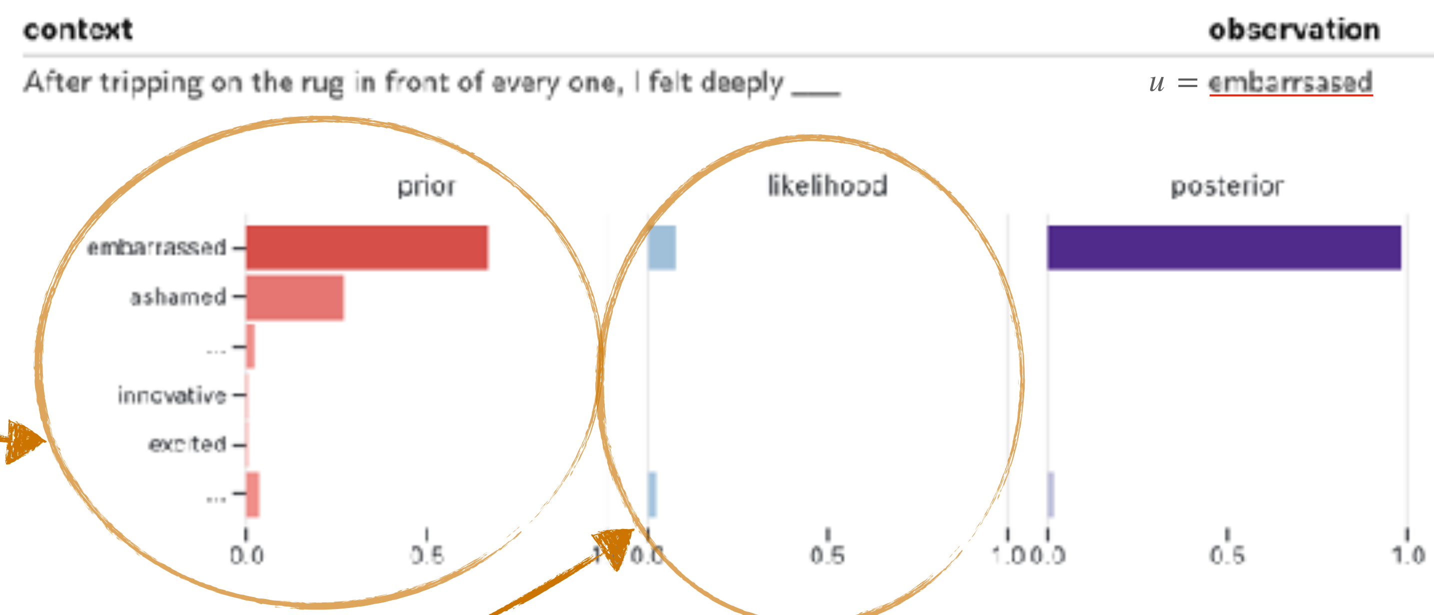
estimating KL and surprisal in noisy channel



generative model:



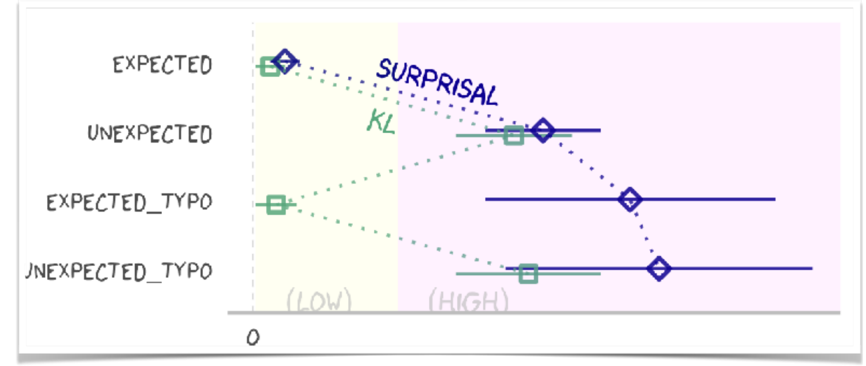
- **prior** over intended words
  $p(z \mid \text{context})$
  = LLM next-seq distribution
  constrained to wordlist
  $\propto p_{\text{LM}}(\text{context}) \odot \mathbf{1}_{\texttt{wordlist}}$



- **likelihood** of observed string:
  $p(u \mid z)$
  = string-edit distance model
  $p(D_{\text{Lev}} \mid z) \cdot p(u \mid D_{\text{Lev}}, z)$

typos as a case study

estimating KL and surprisal



context After tripping over the rug in front of everyone at the party, she quickly got up, but her cheeks turned red and she felt deeply

	z	prior
_embarrassed	6.5668e-01	
_ashamed	2.6608e-01	
_guilty	1.6075e-02	
_uncomfortable	1.0753e-02	
_shy	7.0945e-03	
...		

observation $w =$ embarrassed (expected)

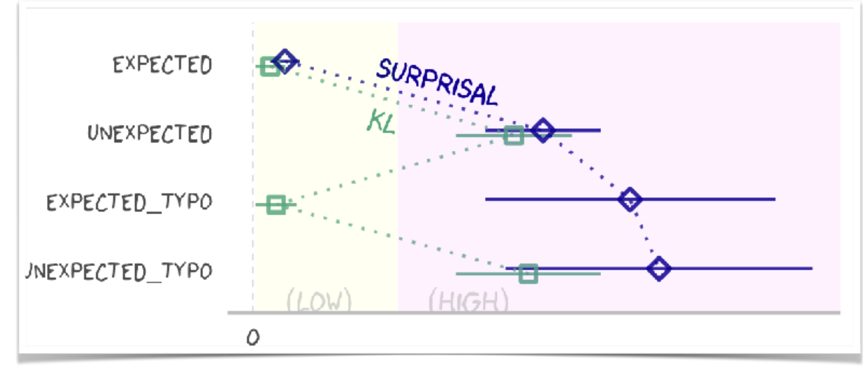
z	prior		likelihood		posterior	
_embarrassed	6.5668e-01		8.9583e-01		1.0000e+00	
_embraced	3.6091e-06		9.4480e-16		5.7964e-21	
_impressed	6.4865e-05		1.6926e-19		1.8663e-23	
_arrested	1.8016e-06		1.6229e-19		4.9701e-25	
...						

prior over intended words
$p(z)$ = LM

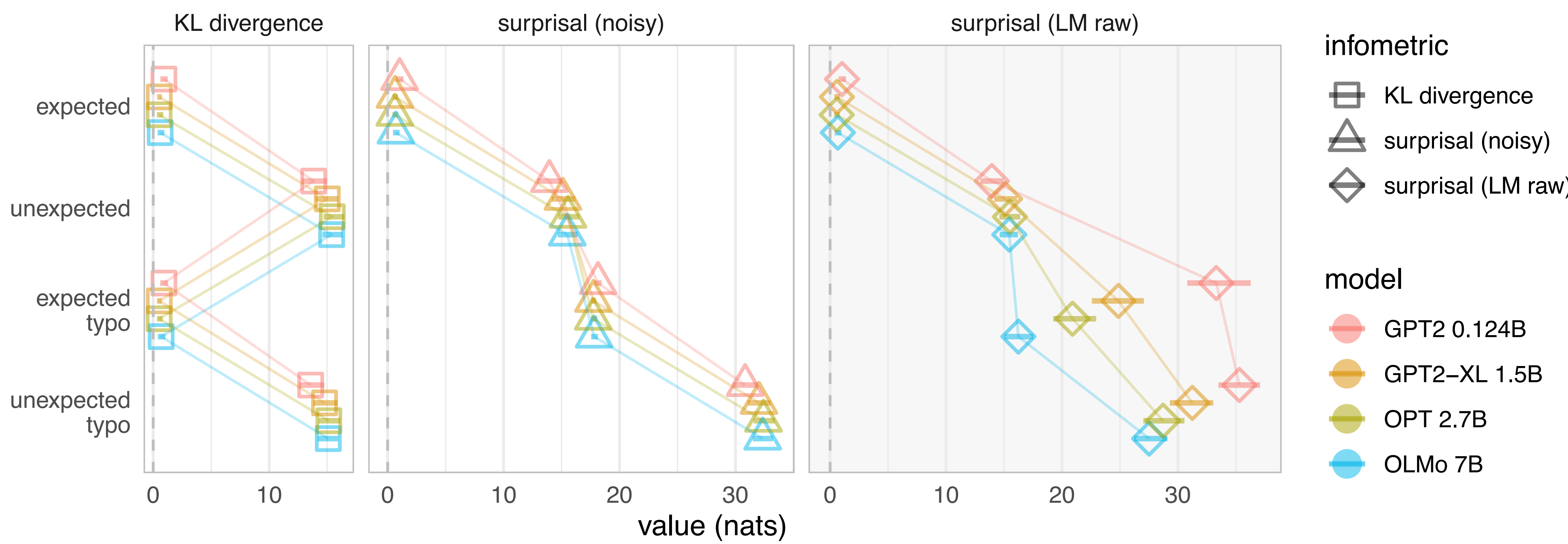
likelihood of observed string:
$p(u \mid z)$ = noisy string model

typos as a case study

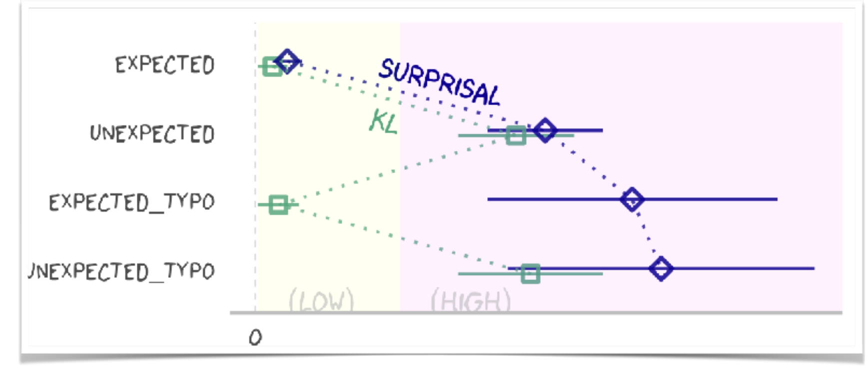
estimating KL and surprisal



Estimated KL divergence and surprisal



typos as a case study



Does surprisal pattern as expected?

Yes. Surprisal is low in **`expected`** condition, but high in others.

Does human RT pattern like surprisal or divergence?

RTs zig-zag, **as update-size predicts**, contra surprisal.

as estimated in our noisy channel model

surprisal theory
(Levy '08)

$$\mathrm{cost}(u) = f(\mathrm{surprisal}(u))$$

update-size theory

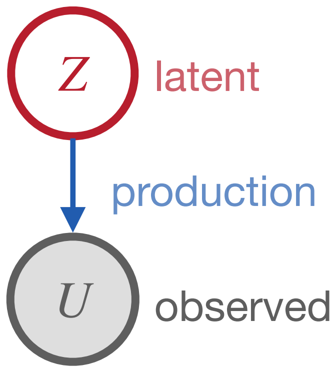
$$\mathrm{cost}(u) = f(\mathrm{D_{KL}}(p_{Z|u} \| p_Z))$$

divergence (belief-update size) **connected to computational complexity of sampling algs.**

$\Longrightarrow$ **motivates** (sampling-based) rational inference algorithms for processing

next steps - better estimates

- for typos
  - more realistic models of typos (using typing statistics)
  - broad-coverage model of KL (not just our materials)
- use **character level LMs** for prior and likelihood models
  - Giulianelli et al. 2024, Vieira et al. 2024
- more broadly: researcher must answer "**what is** Z?"
  - unlike surprisal, requires different models depending on **task**
    - infer intended **words**? **referent**? **sentiment**? etc. (model *task effects*)



thanks

- to you!
- to collaborators: Tim O'Donnell, Peng Qian, Morgan Sonderegger, Steve Piantadosi
- to NSF for SBE postdoc fellowship grant (SMA-2404644)

next steps - applications beyond typos

other places where we think surprisal $\gg D_{\mathrm{KL}}$ (that is, $R \gg 0$):

any (more interesting) constructions where some target region is processed without difficulty despite being very unpredictable

unexpected ways of communicating expected information (thanks to ideas from Alec Marantz)

- synonyms: *This living-room furniture set consists of a table, chair, and couch.* (vs *sofa*)
- epithets: *Boy do I hate that guy John. From the moment the bastard came in the room ....*

grammatical illusions (see e.g., Zhang et al. 2023, 2024)

- Moses illusions: *In the biblical story of the Ark, how many animals of each kind did Moses take with him?*
- agreement attraction: *The key to all the cabinets are on the table.*
- NPI illusions: *The bills that no senator voted for will ever become law.*
- depth-charge illusions: *No head injury is too trivial to ignore.*

malapropisms

- *Sure, if I reprehend (apprehend) anything in this world it is the use of my oracular (vernacular) tongue, and a nice derangement (arrangement) of epitaphs (epithets)!* (Sheridan, 1775)

multilingual codeswitching

- *"Veux-tu rentrer dans ma bubble?"*