processing cost as belief divergence

Jacob Hoover Vigly

MIT BCS
Cog Lunch
29 October 2024

sentence processing

how do we understand what a sentence means?

meaning $z =$

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

- sentence unfolds word by word: $\mathbf{u} = u_1, u_2, \ldots$
- with each word, refine guess about the meaning, z

sentence processing

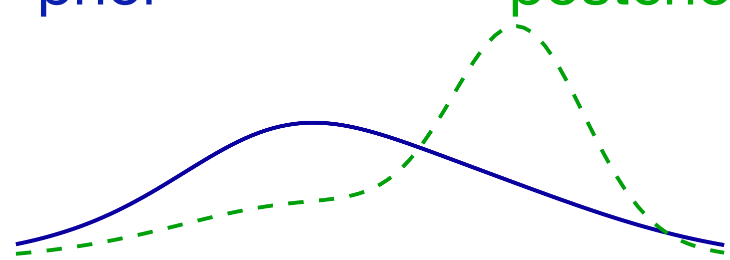
iterative inference problem

$z =$

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

- observe otterance word by word: $\mathbf{u} = u_1, u_2, \ldots$
- with each word, **update beliefs** about the meaning, z

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

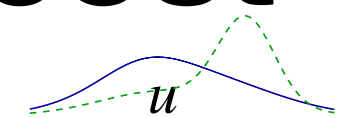


How? …with what processing algorithm?

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

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:

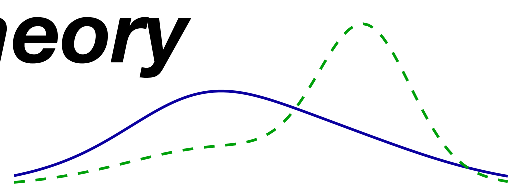
surprisal theory
(Hale '01, Levy '08)

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

precise description of phenomenon, … but what algorithm?

- refocus idea: difficult = big update (large *divergence*)

divergence 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 $\mathrm{D_{KL}} = \mathrm{surprisal}$ (i.e., extra term is zero) ⟵ Let's focus on this one

(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**.
- bigger update = more difficult

common algorithms don't scale in surprisal / divergence

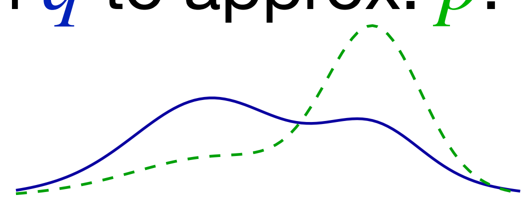
what kind of algorithm *does*?

those that somehow prioritize more probable hypotheses:

- **sampling algorithms**

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

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



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

$$= f(\text{surprisal})$$ assumption (a)

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]

when surprisal ≠ divergence

surprisal theory

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

divergence theory

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

recall motivation: surprisal as measure of size of belief update

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

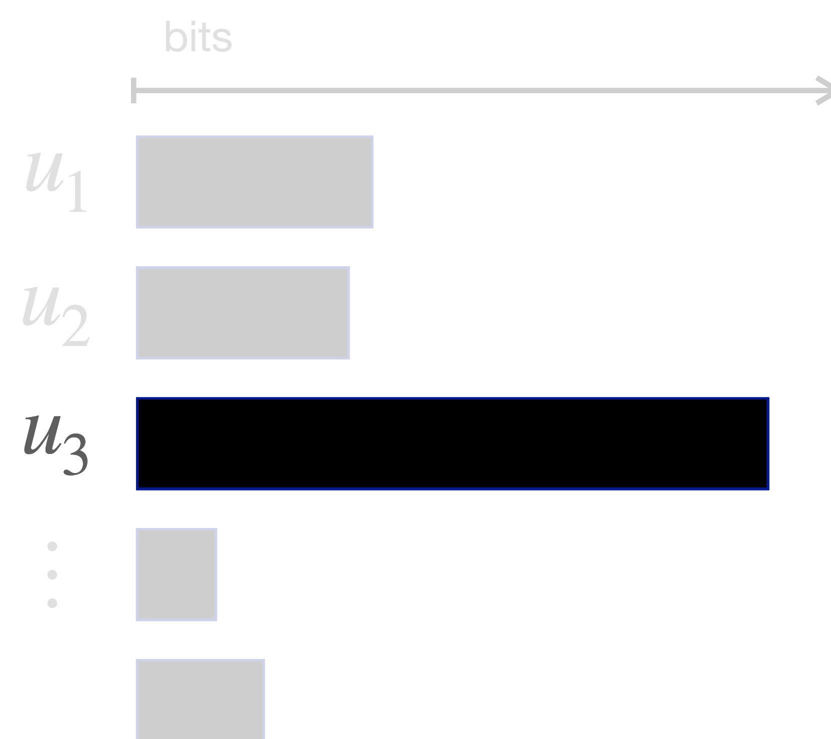
when surprisal < divergence

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

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



When unpredictable does not mean difficult
(WIP with Peng Qian, Morgan Sonderegger, Tim O'Donnell)

typos as a case study

bit.ly/KLdemo

interactive

typos as a case study

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

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

(even with correct 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*

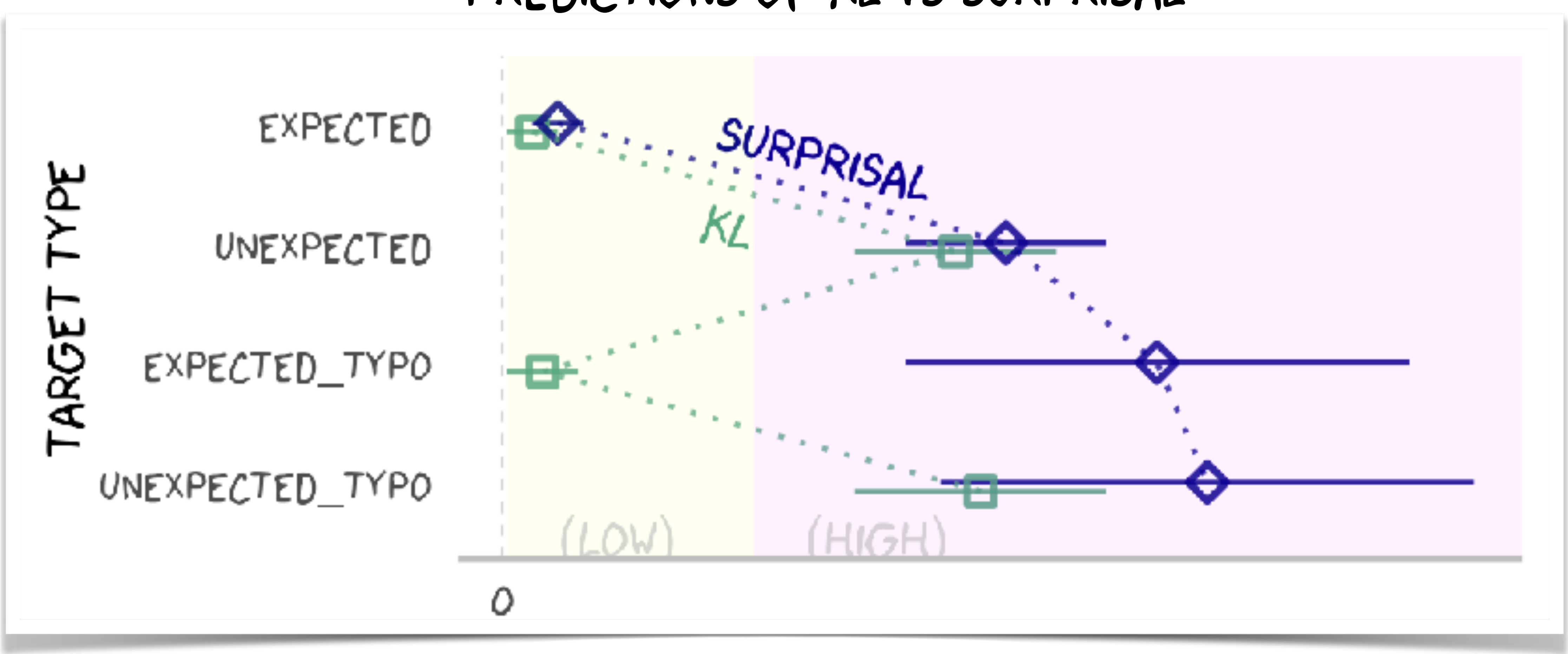
Self-paced reading time study:

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

Fit mixed-effect regression models:

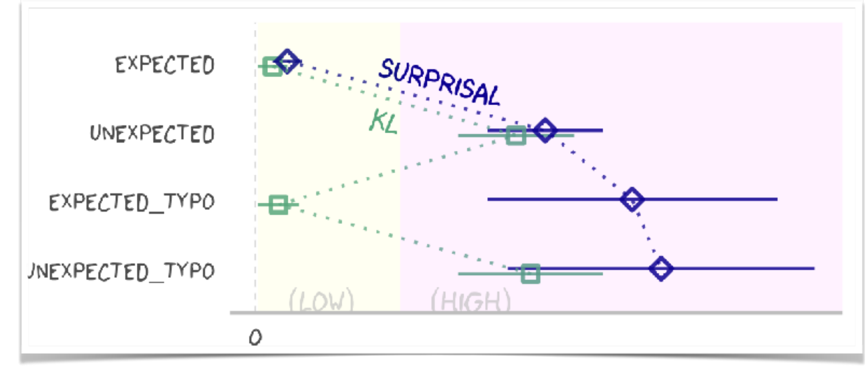
- predict human RT
- predict LLM surprisal (separately)
  - surprisals from collection of LLMs

PREDICTIONS OF KL VS SURPRISAL



When unpredictable does not mean difficult
(WIP with Peng Qian, Morgan Sonderegger, Tim O'Donnell)

typos as a case study



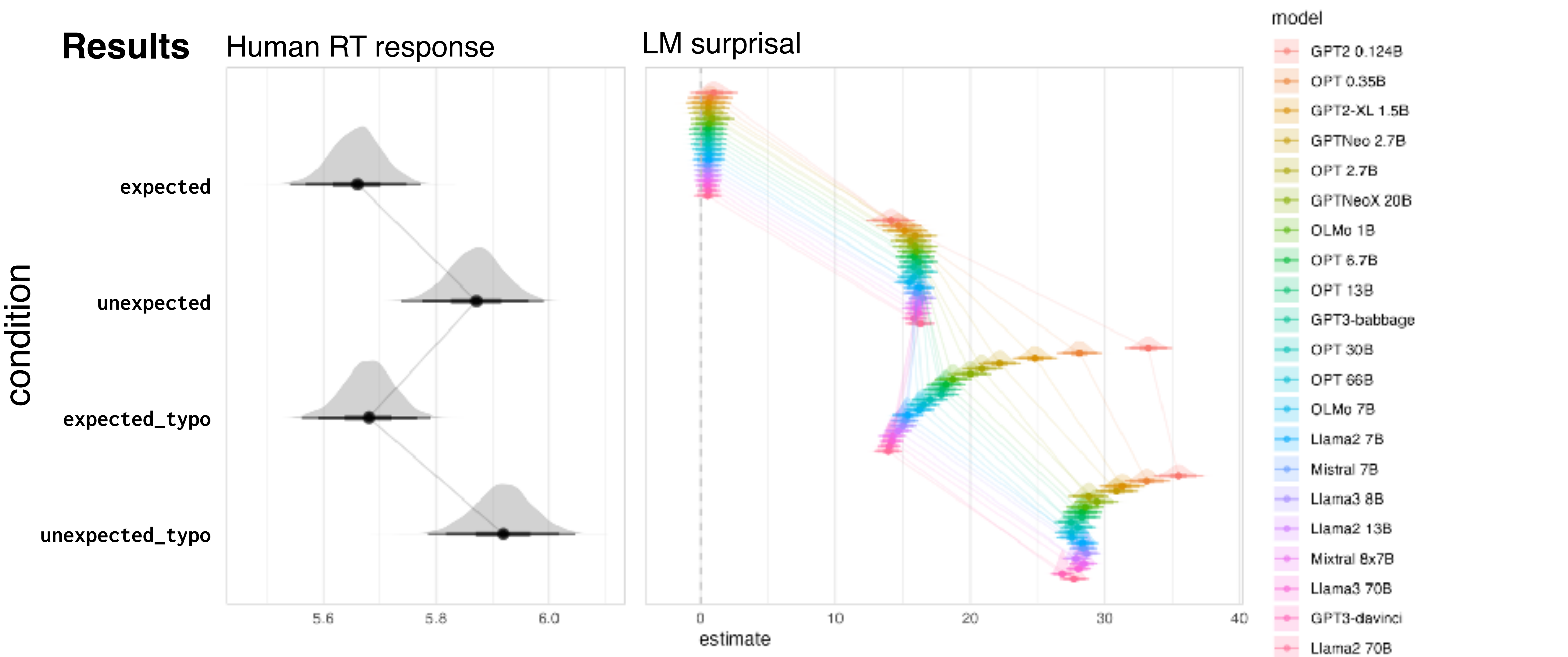
Does surprisal pattern as expected?

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

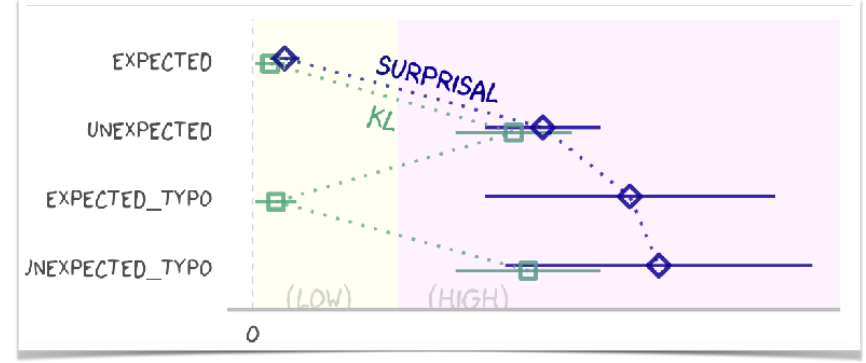
Does human RT pattern like surprisal or divergence?

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

Results



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 would predict**, contra surprisal.

surprisal theory
(Levy '08)

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

divergence theory

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

divergence (information gain) can be **directly connected to sampling complexity**

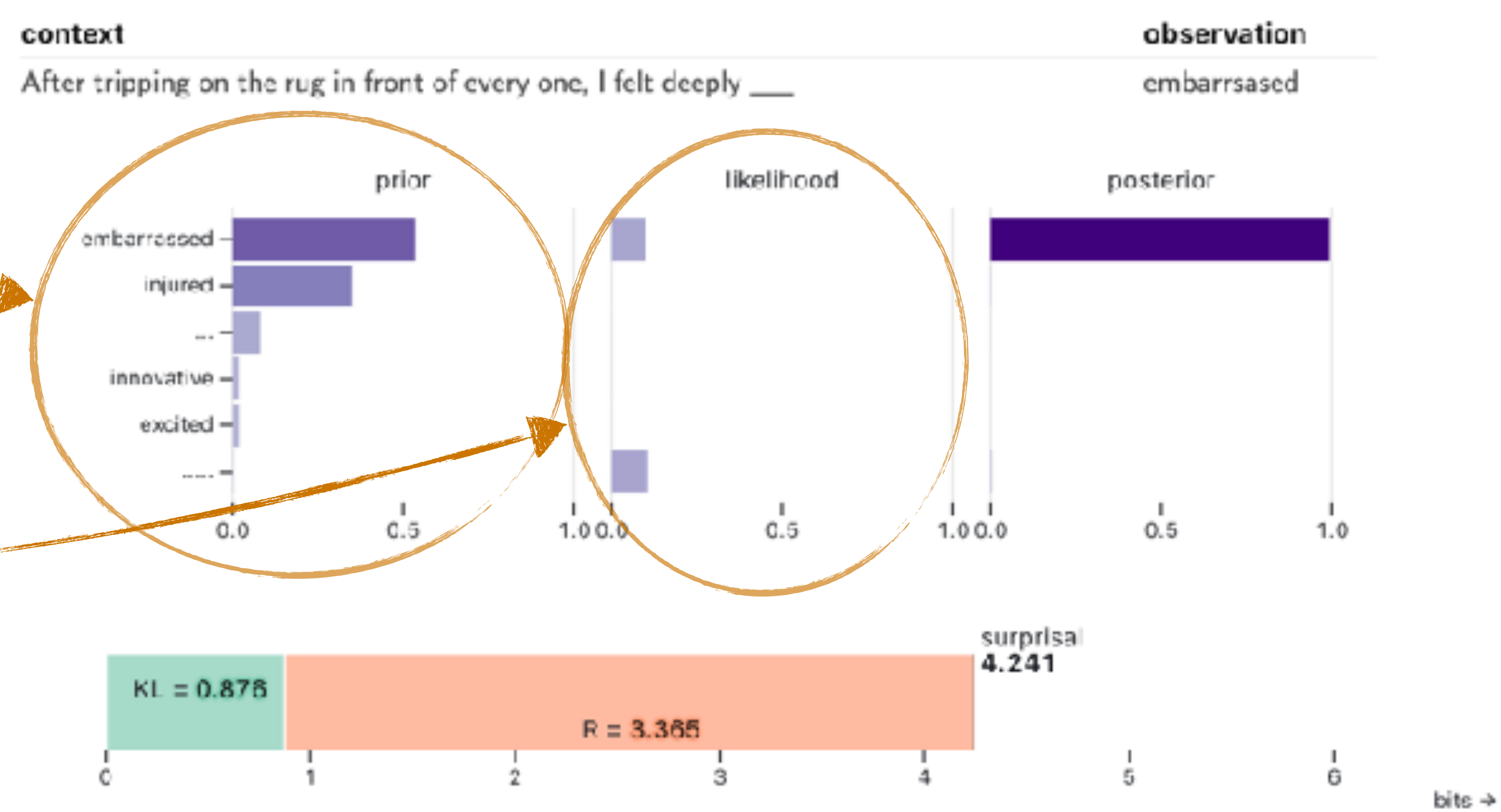
$\Longrightarrow$ **motivates** sampling-based inference algorithms for processing

next steps - getting divergence estimates

- compute estimates of KL for typo study
  - for this, need model of **prior** distribution and **likelihood** function

prior over intended words: pretrained LLM (constrained to use dictionary words)

likelihood of observed string: use production data (e.g. from TypeRacer) to model $\lambda z\,.\,p(u \mid z)$



next steps - not just 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

(Currently collecting stimuli)

grammatical illusions

- 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?"*

THANK YOU!

concluding

how to explain processing cost?

phenomenon: for humans, **unexpected words take longer**

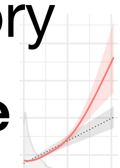
- **but** what type of processing algorithm that can explain this?

I propose **algorithms involving sampling are promising**

- complexity intrinsically scales in statistical properties of input
- $\Longrightarrow$ reframe surprisal theory as **divergence theory:** $\text{cost} = f(D_{\text{KL}})$

I challenge two assumptions of traditional surprisal theory

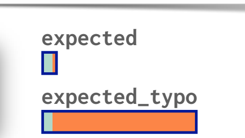
- 1: evidence link is **superlinear**, with increasing **variance**
- 2: evidence that situations exist where **KL ≠ surprisal**



The Plausibility of Sampling as an Algorithmic Theory of Sentence Proces

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

When unpredictable does not mean difficult
(WIP with Peng Qian, Morgan Sonderegger, Tim O'Donnell)



Suggests sampling can explain how cost is related to distributional information

next steps:

- implement direct estimators of KL divergence for use as predictors of processing cost
- develop **sampling-based inference algorithms for parsing**

concluding

next steps

- Look at semantic/syntactic phenomena where potentially KL theory and surprisal theory differ (e.g. grammatical illusions)
- **build SMC model** of comprehension in noisy channel
  - Particle filter (pretrained LM + noise model) or potentially with GenParse
  - Good setting to explore algs with adaptive number of particles
- Develop KL theory from a proposal distribution
  - i.e. what if we use **something smarter than the prior, given observation**?

cost hypothesis	cost(u) =	assumptions
information gain	$f(\ \text{information-gain}(u)\ )$	processing cost scales with information gain
KL from proposal	$f(\ \underbrace{D_{KL}(p_{Z\mid u} \| q_{Z;u})}_{\text{surprisal}(u) - [R(u)+D(u)]}\ )$	& information gain quantified as KL between proposal q and posterior
KL from prior	$f(\ \underbrace{D_{KL}(p_{Z\mid u} \| p_Z)}_{\text{surprisal}(u) - R(u)}\ )$	& proposal q is the prior
general surprisal	$f(\ \text{surprisal}(u)\ )$	& $R(u)$ is zero (binary likelihood)
standard surprisal	$\beta\ \text{surprisal}(u)$	& Linking function f is linear