**Sampling-Based Assumptions Simplify Modeling of Speeded Cloze Response Times**

Sathvik Nair[1], Philip Resnik[1], Colin Phillips[2,1] *University of Maryland[1], Oxford University[2]*

**Motivation:** Lateral inhibition between competing candidates is a well motivated mechanism in models of single word recognition and production; mapping from concepts or wordforms to lexical items. The status of lateral inhibition is more contested in models of prediction; mapping from contexts to lexical items. For prediction in comprehension, experimental [2, 3] and modeling [4] evidence argue against a role for lateral inhibition. For prediction in production, such as in speeded cloze paradigms, the evidence is mixed [5-7]. Staub et al. [5] argue that it is not needed, but others [6, 7] argue that it is needed in order to successfully fit models to human RT data. Importantly, these models are based on the assumption of underlying activation times that follow a Gaussian distribution. Here, we show that a model of context-to-word mappings that is understood as a sampling process [8] yields estimates that capture the range of human RTs without assuming lateral inhibition. This is because the sampling process implies a geometric distribution of lexical access times, rather than a Gaussian distribution. Furthermore, the sampling mechanism also makes it feasible to model human RTs using estimates derived from real experimental items rather than constructed data used in previous models.

In their original model, Staub et al. simulate a cloze task by sampling from distributions of hypothesized RTs (*finishing time distributions)* for 10 completions that are activated by the context [5]. The word that gets produced is then the one with the smallest finishing time out of the 10 samples. However, their modeling results do not capture the strong association between cloze probabilities and RTs in their empirical data (Figure 1), and cannot explain why low-cloze items get produced. For the model to account for these concerns, Nakamura & Phillips incorporated an assumption of lateral inhibition [6]. Both models assume Gaussian distributions for activation and finishing times, without justification. A further empirical limitation is that they both work over hand-constructed data (Figure 2a), rather than values derived from the experimental materials.

**Method:** We remedy these limitations by adjusting the form of the finishing time distributions. To do so, we extend recent claims linking language processing to rational probabilistic inference [8,9]. In particular, Hoover et al. [8] claim inference can be reflected by iteratively sampling from a prior distribution over possible candidates. The length of the sampling process reflects cognitive effort to process a given meaning. Under their simple guessing algorithm, the number of times needed to sample a candidate meaning (here, operationalized as a word) follows a geometric distribution parametrized by its contextual probability. Sampling a candidate meaning could also be analogous to reaching an activation threshold; Nakamura & Phillips [6] also sample from arbitrary Gaussian distributions over activations and convert iterations into RTs.

We propose cloze RTs reflect a process of sampling multiple preactivated candidates, and picking the one that finishes first. As a proof of concept for this view, we make the candidates' finishing time distributions geometric [8], parametrized by their probability under the language model GPT-2 [10] (Figure 2b). For each context, we assume the candidates are completions provided in the experiment, and simulate 40 (participants in [5]) trials. The winning number of iterations is then translated to an RT using Nakamura & Phillips' transformation, which assumes one iteration is 25 ms and adds 350 ms for post-lexical processes [11].

**Results & Discussion:** The relationship between cloze probability and RT in our simulations (Figure 3) qualitatively corresponds better with the empirical data, relative to Staub et al. 's simulations (Figure 1). Making the finishing time distributions geometric offers us an elegant way to explain the longer finishing times for low-cloze items without adding additional assumptions of inhibition. If a low-cloze item wins the race, this simply means that it was likely sampled from the tail of its finishing time distribution (Figure 2b). Based on our proof of concept modeling work, explicitly implementing sampling-based models [12] and connecting them with predictive coding models [4] is a promising future direction to model production as well as comprehension, as both approaches reflect approximate inference under resource constraints [4, 8, 12].

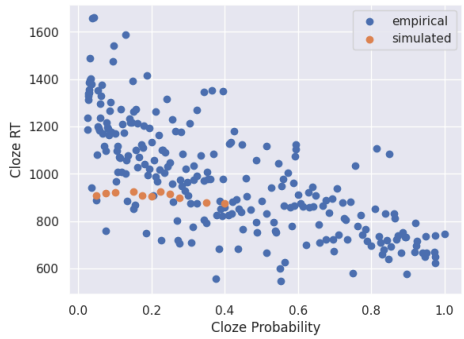
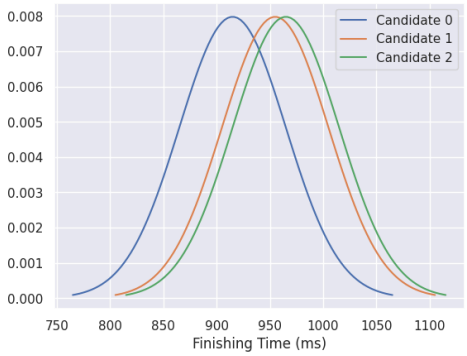


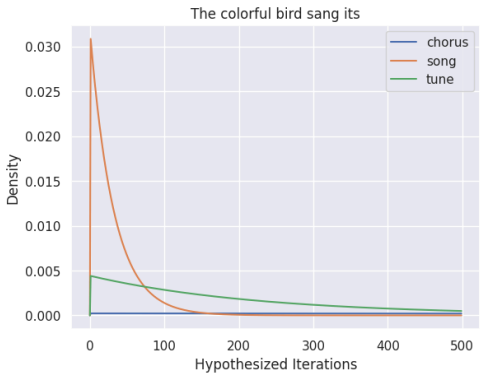

**Figure 1:** Comparison between cloze RTs from Staub et al.'s original model and their experimental data

| Word | LM | Race | Cloze |
|---|---|---|---|
| Chorus | 0.000222 | 0.05 | 0.027 |
| Song | 0.0308 | 0.9 | 0.891 |
| Tune | 0.00439 | 0.125 | 0.081 |

**Table 1:** Probabilities for completions to the context *The colorful bird sang its…*




**2a.**




**2b.**

**Figure 2:** Finishing time distributions, with Gaussian distributions over arbitrary completions from Staub et al.'s original model (2a) and geometric distributions for experimental completions to a context from their experiment, parametrized by language model probability (2b).

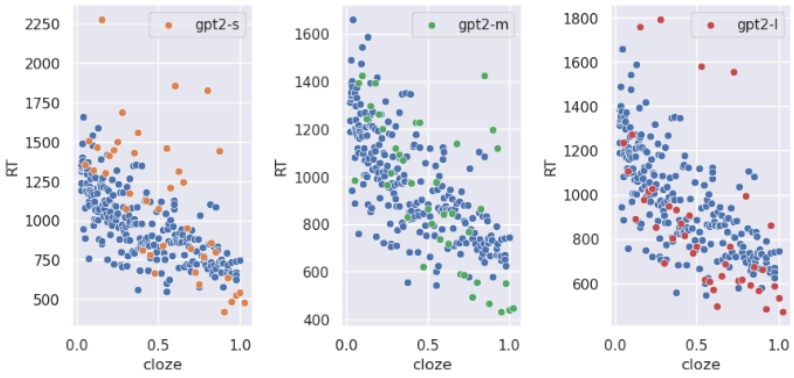



**Figure 3:** Comparison between empirical results (blue) and simulated results under small, medium, and large variants of GPT-2 (left to right). Results are aggregated over cloze probability following Nakamura & Phillips.

**References:** [1] Chen & Mirman (*Psych. Review*, 2012) [2] Laszlo & Federmeier *(JML 2009)* [3] Frisson, Harvey & Staub (*JML 2017)* [4] Nour Eddine, Brothers, Spratling, Wang & Kuperberg, (*Cognition, 2024)* [5] Staub, Grant, Astheimer & Cohen (*JML, 2015)* [6] Nakamura & Phillips (*HSP*, 2022) [7] Ness & Meltzer-Asscher *(Cognition, 2021)* [8] Hoover, Sonderegger, Piantadosi & O'Donnell (*Open Mind,* 2023) [9] Shain, Meister, Pimentel, Cotterrell & Levy (*PNAS*, 2024) [10] Radford, Wu, Child, Amodei & Sutskever (*OpenAI Technical Report,* 2019) [11] Indefrey & Levelt (*Cognition, 2004)* [12] Clark, Vigly, Gibson & Levy, (*EMNLP, 2025)*