Position: Don't Use the CLT in LLM Evals With Fewer Than a Few Hundred Datapoints

(...it's *really* easy to do *a lot* better!)

Sam Bowyer [1] Laurence Aitchison [1,*] Desi R. Ivanova [2,*] [1]University of Bristol [2]University of Oxford *Equal Contribution

1. Failures of the Central Limit Theorem

Central Limit Theorem: If $X_1, \dots, X_N$ are IID r.v.s with mean $\mu \in \mathbb{R}$ and finite variance σ^2, then with sample mean $\hat{\mu} = \frac{1}{N}\sum_{i=1}^N X_i$,

$$\sqrt{N}(\hat{\mu} - \mu) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2\right) \text{ as } N \to \infty.$$

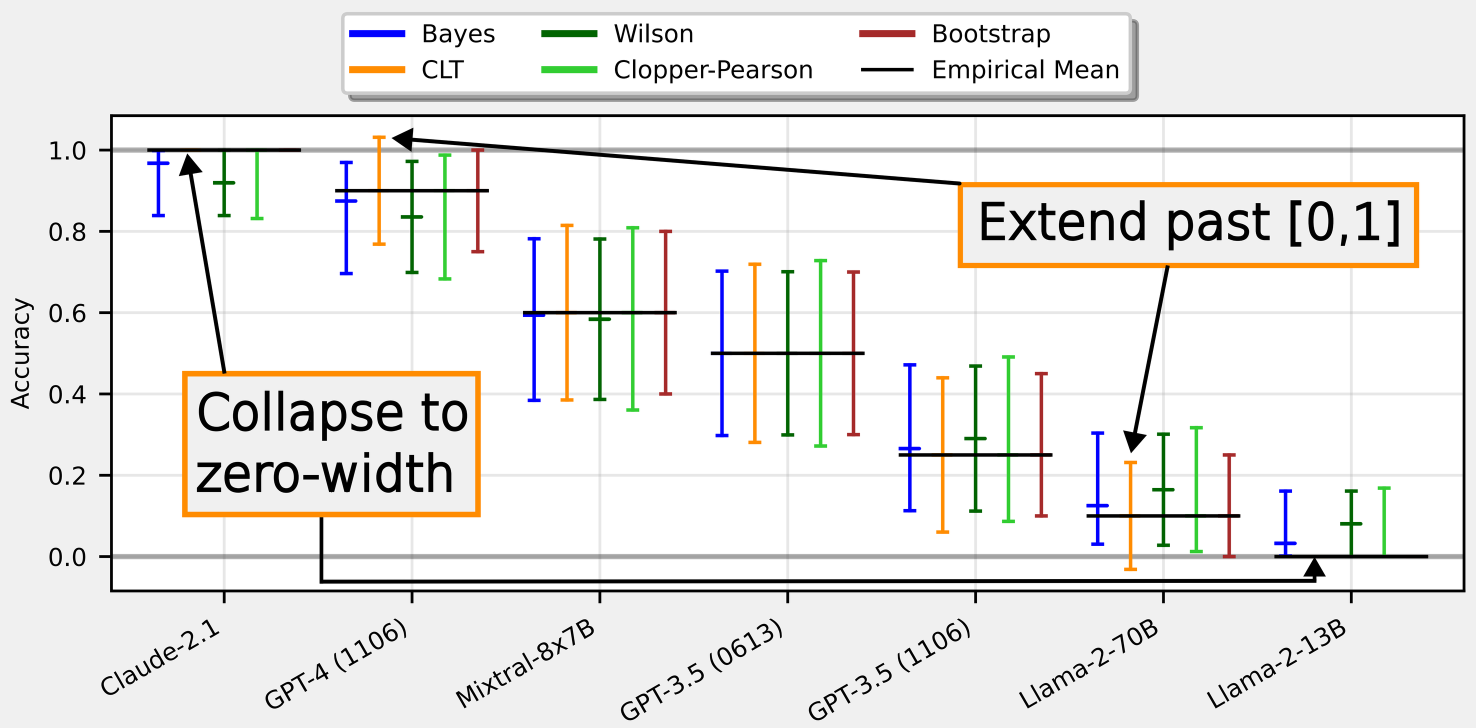
CLT-based confidence interval on binary data

$$y_i = \begin{cases} 0 & \text{question } i \text{ answered incorrectly} \\ 1 & \text{question } i \text{ answered correctly} \end{cases}$$

for $i = 1, \dots, N$, with standard error $SE(\bar{y}) = \sqrt{\bar{y}(1-\bar{y})/N}$:

$$\text{CI}_{1-\alpha} = \bar{y} \pm z_{\alpha/2}\text{SE}(\bar{y}),$$

where $z_{\alpha/2}$ is the $(1-\alpha/2)$-th quantile of $\mathcal{N}(0, 1)$.



As LLM capabilities improve, it's becoming more common to run small-N benchmarks, e.g. the Langchain Typewriter tool-use benchmark shown above ($N = 20$).

2. Simulation Setup

1. **Synthetic datasets:** Draw probability of success
$$\theta \sim \text{Uniform}[0, 1],$$
then draw N IID samples $y_i \sim \text{Ber}(\theta)$.
2. **Construct intervals with various confidence levels**
$$1 - \alpha \in [0.8, 0.995].$$
3. **Repeat** the above many times, and compare different interval-construction methods via
   - **Coverage** (proportion of intervals that contain true value of θ; should equal $1 - \alpha$)
   - **Interval width**

3. Alternative Interval Construction

Bayesian Beta-Bernoulli credible interval — uniform prior over probability of success θ:

$$\theta \sim \text{Beta}(1,1) = \text{Uniform}[0,1]$$
$$y_i \sim \text{Bernoulli}(\theta) \text{ for } i = 1, \dots N$$

Use quantiles of closed-form posterior to construct $1 - \alpha$ CIs:

$$\theta \mid y_{1:N} \sim \text{Beta}\left(1 + \sum_{i=1}^N y_i, 1 + N - \sum_{i=1}^N y_i\right)$$

Beta-Bernoulli Bayesian Credible Interval
```python
posterior = scipy.stats.beta(1 + sum(y), 1 + N - sum(y))
bayes_ci  = posterior.interval(confidence=0.95)
```

Wilson Score confidence interval — based on binomial distribution:

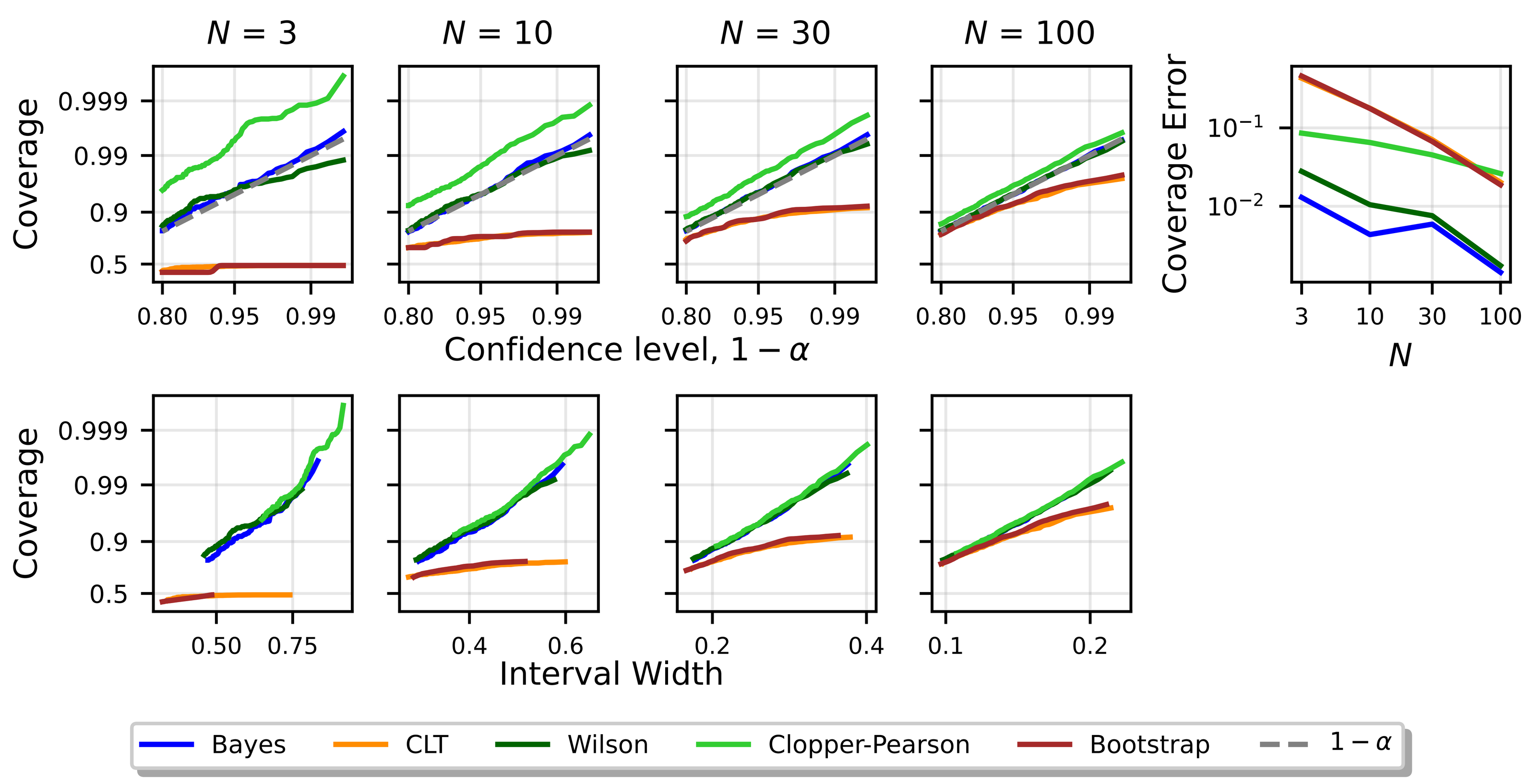
$$\text{CI}_{1-\alpha,\text{Wilson}} = \frac{\hat{\theta} + \frac{z_{\alpha/2}^2}{2N}}{1 + \frac{z_{\alpha/2}^2}{N}} \pm \frac{\frac{z_{\alpha/2}}{2N}}{1 + \frac{z_{\alpha/2}^2}{N}} \sqrt{4N\hat{\theta}(1-\hat{\theta}) + z_{\alpha/2}^2}.$$

Wilson Score Confidence Interval
```python
result = scipy.stats.binomtest(k=sum(y), n=N)
wilson_ci = result.proportion_ci("wilson", 0.95)
```

Clopper-Pearson confidence interval — 'exact' method (will never under-cover). Includes all possible values of θ_0 that are *not* possible to reject at the $1 - \alpha$ level with the null hypothesis $H_0 : \theta = \theta_0$.

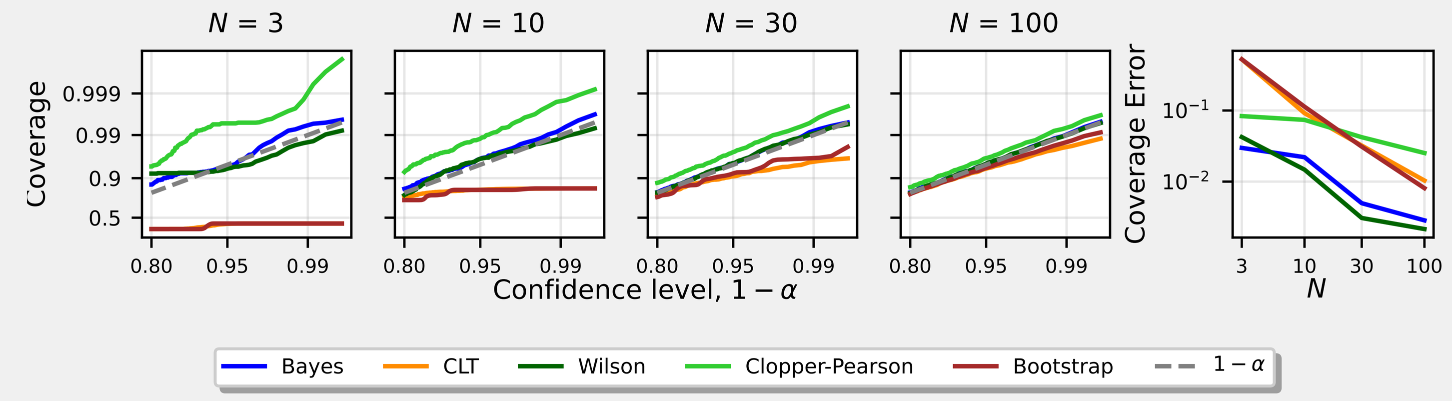
Clopper-Pearson Confidence Interval
```python
 result = scipy.stats.binomtest(k=sum(y), n=N)
 clop_ci = result.proportion_ci("exact", 0.95)
```

4. Results



5. Other Settings

- **Mismatched Prior** — Different prior for data simulation e.g. $\theta \sim \text{Beta}(100, 20)$, $\mathbb{E}[\theta] = 0.83$, $\text{Var}[\theta] = 0.034^2$.

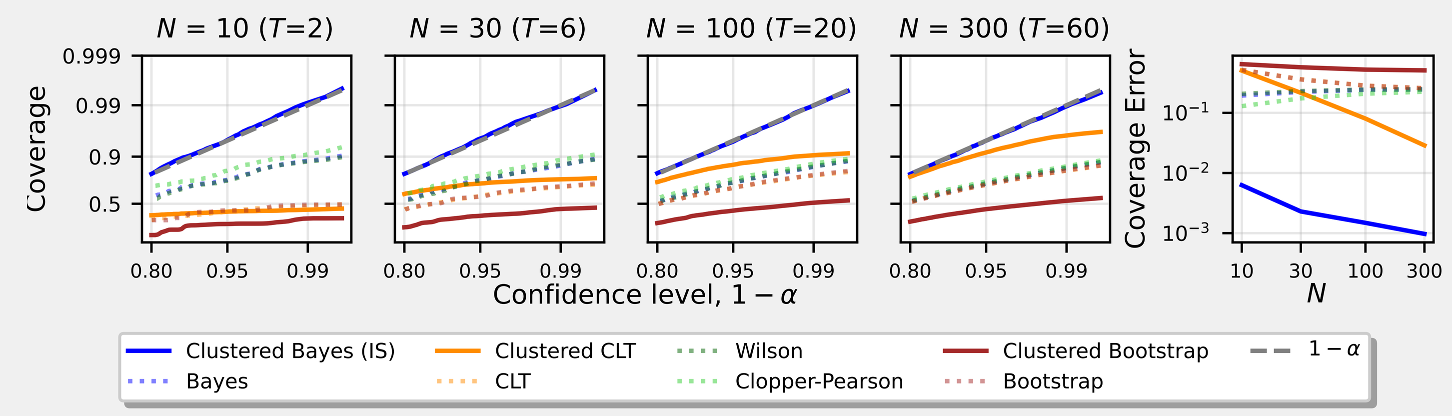


- **Clustered Questions** — T tasks, each with N_t IID questions. Clustered CLT uses a different standard error formulation:

$$\text{SE}_{\text{clust.}} = \sqrt{\text{SE}_{\text{CLT}}^2 + \frac{1}{N^2}\sum_{t=1}^{T}\sum_{i=1}^{N_t}\sum_{j \neq i}(y_{i,t} - \bar{y})(y_{j,t} - \bar{y})}.$$

Bayesian credible interval found via importance sampling:

$$\theta \sim \text{Beta}(1,1), \qquad d \sim \text{Gamma}(1,1),$$
$$\theta_t \sim \text{Beta}(d\theta, d(1-\theta)), \qquad y_{i,t} \sim \text{Bernoulli}(\theta_t).$$



- **Independent Comparisons** — Compare θ_A and θ_B for two different models, with access *only* to $N_A, N_B, \bar{y}_A$, and $\bar{y}_B$.
- **Paired Comparisons** — Compare θ_A and θ_B for two different models, each with the same N IID questions and access to question-level successes $\{y_{A;i}\}_{i=1}^N$ and $\{y_{B;i}\}_{i=1}^N$.
- **Metrics that aren't simple averages** — e.g. F1 score.

We construct Bayesian credible intervals that outperform CLT-based intervals in all settings (available in `bayes_evals`).

6. Recommendations

IID setting: use **Beta-Bernoulli** credible intervals or **Wilson score** confidence intervals via `scipy` or equivalent.

Other settings: use **Bayesian credible intervals** as implemented in our simple package `bayes_evals`.