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

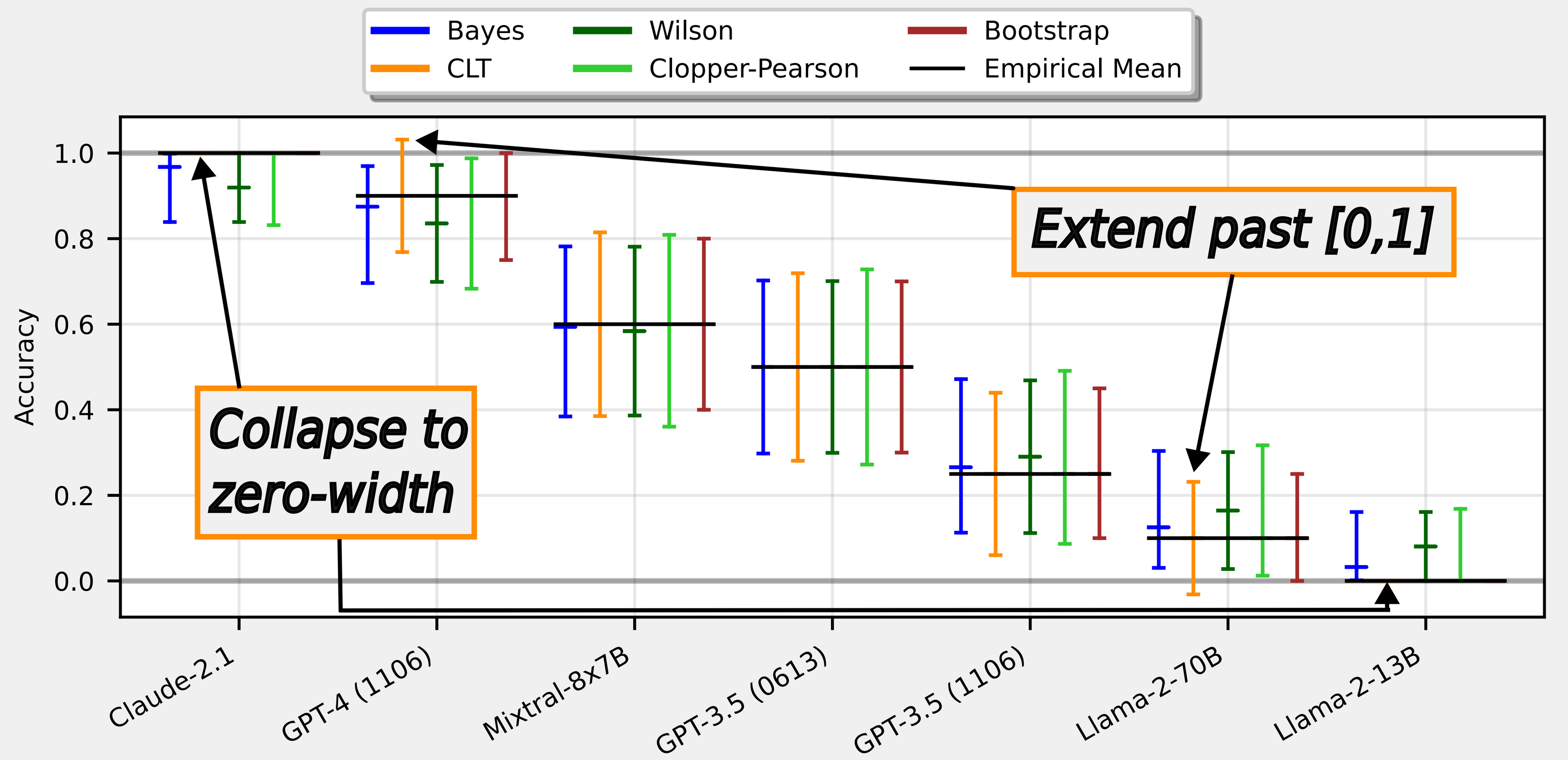
Central Limit Theorem: If $X_1, \dots, X_N$ are IID r.v.s with mean $\mu \in \mathbb{R}$ and finite variance σ^2, then with sample mean $\hat{\mu} = \frac{1}{N}\sum_{i=1}^N X_i$,

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

CLT-based confidence interval on binary data $y_i \in \{0, 1\}$ (incorrect/correct) for $i = 1, \dots, N$:

$$\mathrm{CI}_{1-\alpha} = \bar{y} \pm z_{\alpha/2}\sqrt{\bar{y}(1-\bar{y})/N},$$

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, such as the Langchain Typewriter tool-use benchmark shown above ($N = 20$).

2. Simulation Setup

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

3. Alternative Interval Construction

Bayesian Beta-Bernoulli credible interval — uniform prior on θ:

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

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

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

Beta-Bernoulli Bayesian Credible Interval

```
posterior = scipy.stats.beta(1 + sum(y), 1 + N - sum(y))
bayes_ci  = posterior.interval(confidence=0.95)
```

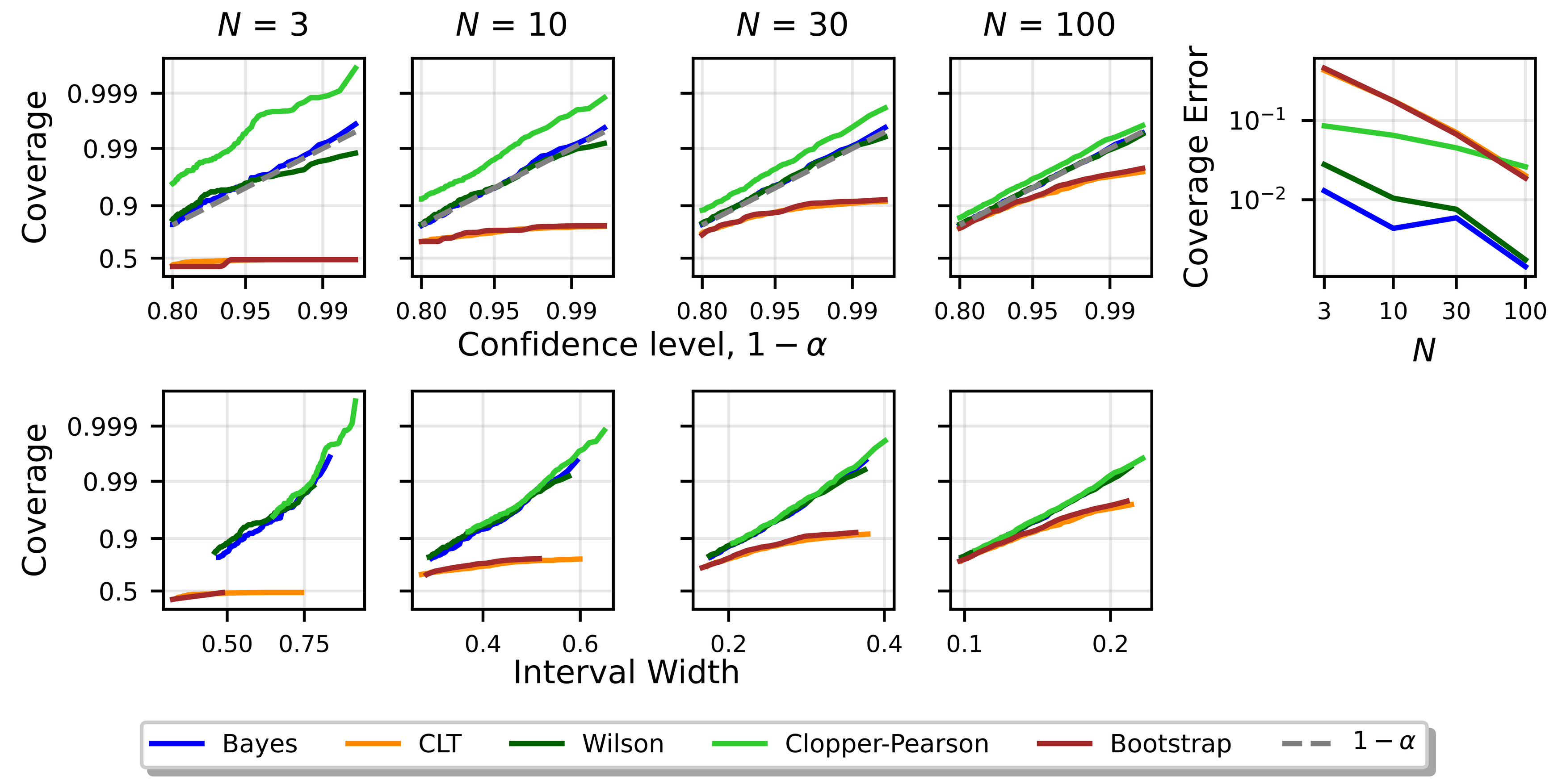
Wilson-Score confidence interval — based on binomial distribution:

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

Wilson-Score Confidence Interval

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

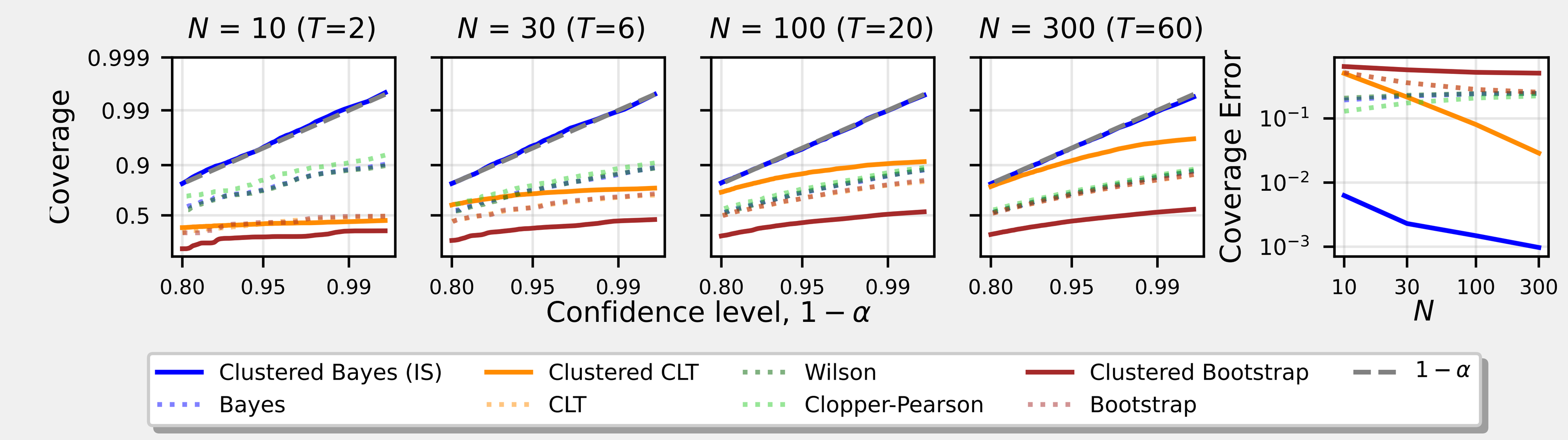
4. Results



5. Other Settings

- **Clustered Questions** — instead of N IID questions, we have T tasks, each with N_t IID questions. Bayesian model:

$$d \sim \mathrm{Gamma}(1, 1), \quad \theta \sim \mathrm{Beta}(1, 1),$$
$$\theta_t \sim \mathrm{Beta}(d\theta, d(1-\theta)), \quad y_{i,t} \sim \mathrm{Bernoulli}(\theta_t)$$



- **Independent Comparisons** — Compare θ_A and θ_B for two different models, with access *only* to $N_A, N_B, \hat{\theta}_A$, and $\hat{\theta}_B$.
- **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 of binary results** — e.g. F1 score (harmonic mean of precision and recall).

We construct Bayesian credible intervals that outperform CLT-based intervals in all settings; implemented in `bayes_evals`.

6. Recommendations

IID setting: use Bayesian 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`.