**Feedback on Figure 7**

*Claim*: Figure 7 shows a 100-point Elo gain after testing 20 pre-release models.

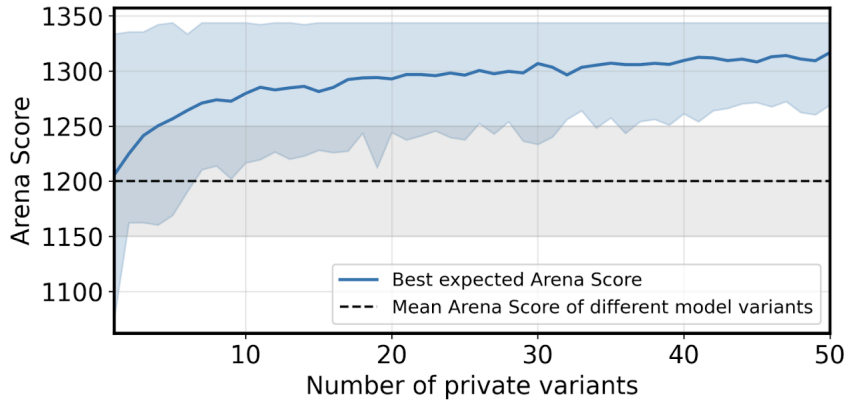
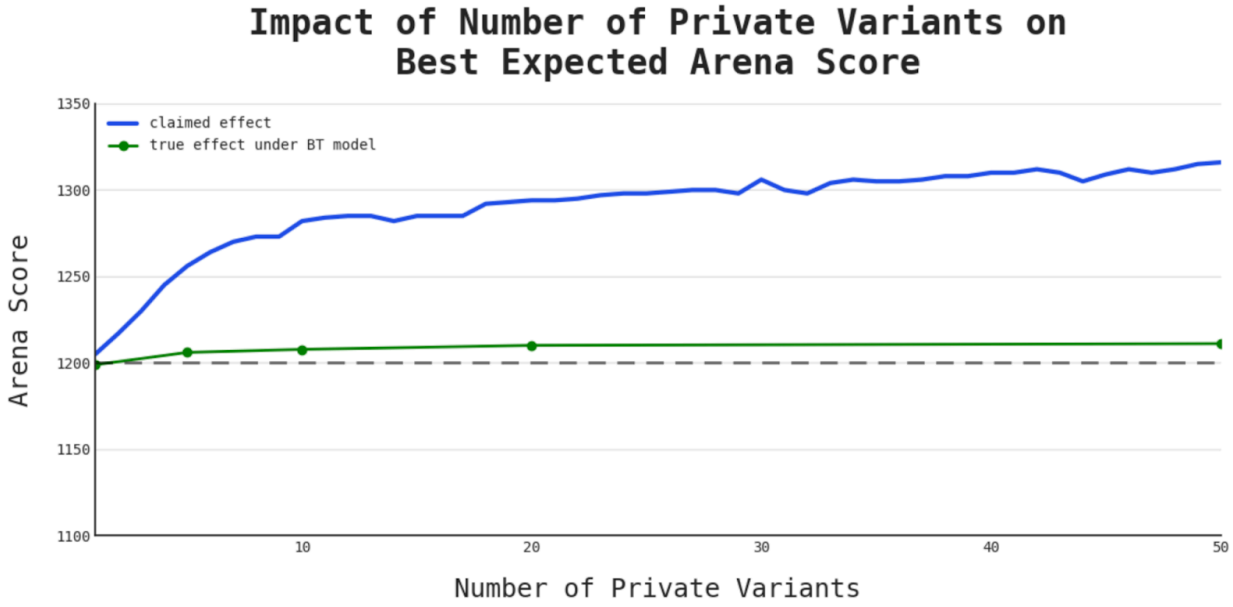



Figure 7: **Impact of the number of private variants tested on the best Expected Arena Score.** We simulate a family of model variants with a latent average Arena Score of 1200. As we progressively increase the number of private variants tested—and subsequently discover their corresponding Arena Scores—the probability of selecting models from the higher end of the performance distribution also rises. This enables the provider to effectively identify the model with the highest score.

*Feedback*: We simulated multiple identical submissions to Chatbot Arena and found negligible differences between their Arena Scores after testing 50 variants. The effect is an order of magnitude less than claimed.




*Short explanation*:

The simulation in Figure 7 is incorrect on several counts.

- It does not use the Bradley-Terry model, and only works by sampling Gaussians. In words, Figure 7 plots the expected maximum of Gaussians with mean 1200 and variance V—this is not a correct model for selection bias in Chatbot Arena. The result of the experiment is *entirely* driven by the choice of V.

- The selection bias induced by selecting the max BT coefficient has an analytic scaling law. It is known to be $C\sqrt{\frac{\log M}{n}}$ , for some constant $C$, where n is the number of battles. In Chatbot Arena, n is at least 3000, and M is at most 50, making this number tiny.

- Chatbot Arena continually evaluates models on fresh new user queries. Because of this, the selection bias is asymptotically zero. Empirically, even a small time-period of continual testing, the bias is negligible.

- The simulation conflates the model choice (e.g. different fine-tunes of a model have different true scores) and randomness due to finite-sample uncertainty in the estimates of the model's score. The former is not a problem; helping select the best model is part of a benchmark's job! The latter is asymptotically zero. The simulation would argue otherwise, which is false.

Instead, the experiment we ran directly simulates the finite-sample selection bias in the estimation of Bradley-Terry coefficients. It does so by using real data from Chatbot Arena to initialize the simulation, making it a much more accurate representation of the true bias.

Details in the full writeup [PDF link].