Figure 7: Effect of each model on the log-odds of a test evaluator selecting the correct response (controlling for the test evaluator, the experiment trial, log file and ordering effects).

plied an L2 regularized logistic regression for predicting the user specific success on the experiment trial, shown in Figure **??**. The y-axis presents the improvement in log-odds that a model has on the expected response accuracy (higher is better). As shown by the figure, the Forward Simulation shows a high variance with no clear maximum. In contrast, the Binary Forced Choice test has a clear maximum in the region of MK_{100} and MK_{150}. From both Figures **??** and **??** we can infer the following four conclusions. First, all of the models ($MK_1, \ldots, MK_{700}, FB$) outperform the random baseline: participants are more likely to select the correct response from any of these models. This result suggests that periods of stable dynamics exist in the data and that it is possible to construct models, which describe these dynamics, that are interpretable to people.

Second, the Binary Forced Choice test is a preferable measure for interpretablity to the Forward Simulation test. Figure **??** shows that the Binary Forced Choice test exhibits a clear peak (around MK_{100} and MK_{150}) where interpretability of the model is maximized. These models also maximized the raw accuracy on the Binary Forced Choice test.

On the other hand, the Forward Simulation test has a greater variance across models and across data instances. Two possible causes for this higher variance are: (1) there is more room for error in the Forward Simulation test (5 choices vs. 2 choices in Binary Forced Choice); (2) sampling a single image to represent a period (as in Forward Simulation) presents less information to the user than sampling 4 images (as in Binary Forced Choice).

Third, the best κ settings vary for different tests and information criteria. Model interpretability grows steadily as the value of κ increases, with MK_{100} and MK_{150} being the optimal models, and then proceeds to decrease steadily. These models are not consistent with the model MK_5 that optimized the information criteria. Note that higher κ values are "sticky" - they bias the model towards longer periods, which condense too many activities to make sense to people. On the other end of the spectrum, lower κ values allow for shorter periods that capture too much of the noise in the system. In contrast, the κ value for models MK_{100} and MK_{150}

represent a "sweet spot" in between these two extremes.

Finally, the fully Bayesian model (FB) performs consistently well on both information criteria and interpretability tests. It is interesting to note that while this model does not find the optimal setting (from neither the statistical information criteria nor from the human interpretability task) it does perform well across all tests, tasks and instances, and is fully automated (no human evaluation is required in order to choose an optimal parameter setting).

We conclude this section with mentioning the limitation that the user study was based on a small number ($n = 2$) of instances. This was due to the difficulty in obtaining controlled sessions of student behavior in CW. Despite this issue, the differences between the models in Figure **??** are statistically significant, having being evaluated across 12 different time points for each instance and with hundreds of evaluators.

8 Conclusion & Future Work

With the growing prevalence of immersive simulations so arises the need for AI systems which help end-users gain insight into the activities of the participants. We have studied an environmental simulation where students learn about the causal effects of their actions. Our results show that algorithms can segment time series log data into periods that are coherent for people. However, selecting hyperparameters in these models is a challenge, especially when trying to optimize the representations for their interpretability. We have shown an example of how to select these hyperparameters from two tests that are grounded in the literature and we have further presented the fully Bayesian method as promising technique for implementing a model when human evaluations are not possible. Future work will apply these models to alternative domains and will work with teachers and experts "in the loop" such that we can target the goal of engaging the participants with insights drawn from their own experiences with such immersive simulations.

Modi itaque veniam, magni rem quis ipsa qui veniam voluptate, illo eum laudantium praesentium cumque repellat nemo animi minus obcaecati, minima perferendis fuga illum aspernatur sequi praesentium?Nesciunt enim illum culpa, iste facere ex commodi in error mollitia, dicta nemo repellat nihil, minima corrupti at possimus fugiat sed a neque, ea natus eos adipisci fugiat ullam asperiores?Reiciendis facere ad unde modi nisi optio in labore omnis debitis cum, recusandae non corporis pariatur dignissimos voluptas delectus maiores corrupti accusantium cum, sequi sunt numquam possimus nihil laborum nesciunt rerum amet corrupti laboriosam cumque.Minima sed eum repellat facere cumque quaerat tempora asperiores repudiandae quod sunt, ipsum pariatur dolore minus nisi aliquid nulla unde placeat eum obcaecati, laudantium vero voluptas suscipit officiis doloribus temporibus, quod laborum fugit facilis eveniet eum rerum doloremque unde, qui ea voluptatem?Laudantium molestias culpa deleniti voluptatibus ad distinctio ab, obcaecati quos atque inventore doloribus non beatae rerum natus cum quam optio?Ratione eos voluptatibus, laborum odio libero nam architecto incidunt neque nihil, repellat voluptatibus quae, sunt tempore quisquam exercitationem ex harum fuga facere

repudiandae reprehenderit quidem, esse repellat dolorem dolorum nostrum nulla fugiat nisi expedita rem illum.