We would like to thank the reviewers for their insightful reviews. We are glad that the feedback was generally positive, as the reviewers were happy with the motivation and practicality of our work.

The primary weakness that several reviewers brought up was that the methods and analysis were straightforward. We believe that our novelty is in proposing Thompson sampling latent bandit algorithms using offline-learned graphical models (PGMs) as side information, with full regret analysis. Reviewers are correct in noting that our analysis relies on insights made earlier by Russo and Van Roy. However, we are the first to apply it to bandit problems with latent variables, which is very common in many applications e.g., personalized recommender systems. Our near-optimal analysis is very general and we are working on applying it to problems with more complicated and problem-dependent PGMs (e.g., PGMs with latent state transition dynamics or with factored, continuous, latent state structures).

Reviewer #1 *"algorithm ... is not designed keeping short horizons in mind"*: Our algorithms quickly personalize by assuming users can be clustered into a finite set of latent states. When the set is small, identifying which state best describes the user's preferences can happen much quicker than learning the user's preferences from scratch.

Reviewer #2 *"suffers an exploration-exploitation tradeoff"*: You are correct in noting that our algorithm depends on exploration of the offline data to learn good models. If the models are good though, our method achieves much greater sample-efficiency than baselines that ignore the offline data.

"unified analyses cannot cover instance-dependent bounds": We derive Bayes regret bounds, which contain an expectation over possible instances. We agree that instance-dependent regret bounds are more informative, but more difficult to derive and interesting future work. Our view our work as a first step to achieving this.

Reviewer #3 Thank you for your detailed corrections! We will update the paper with your clarifications.

"the available epsilon-bounds are wildly pessimistic": You are correct in noting that our regret bounds require that the offline-learned model has low prediction error ε. Choosing ε using model-learning guarantees will likely lead to an overly conservative `mmUCB` algorithm. However, we view `mmUCB` not as a practical algorithm, but one that can be used to analyze `mmTS`, which is practical and general. The `mmTS` algorithm additionally computes posterior model parameters using online interactions. This means the ε term should not affect its long-term performance, and the regret bound we derived is likely too conservative. Updating our regret bounds to reflect this is a future line of work.

"no attempt to analyze the problem in computational terms": Given a latent variable model, the optimal policy would plan out the entire sequence of actions using the model, akin to solving a special case of a POMDP. It is unclear how to do so feasibly in our setting, even on short horizons. Gittins index will compute an optimal strategy for Bayesian bandits, but we are also unsure how to generalize the method to complex latent models. We instead compare our algorithms to a "post hoc clairvoyant" algorithm that performs at least as well

Relation to value-of-information problem: A key distinction is that prior value-of-information work seems more concerned with discovering the optimal solution, similar to best-arm-identification, whereas our setting deals with regret minimization. We will cite such work by Krause et. al and discuss its differences in our work.

Reviewer #4 *"regret bounds for the misspecified model case ... are linear in n"*: It is true that the performance of our algorithms depend on the quality of the offline-learned model. Past work in offline model-learning e.g. spectral methods, can give guarantees that are $\varepsilon = O(1/\sqrt{n_{\mathsf{offline}}})$ where n_{offline} is the size of the offline dataset. On short horizons $n \leq n_{\mathsf{offline}}$, the contribution of ε to overall regret is small. It is also important to note that our proposed `mmTS` algorithm computes posterior model parameters given online interactions. This means the ε term due to the prior is too conservative. Having a regret bound that reflects this property remains as future work.

Questions about latent states: We consider the latent bandits setting in Maillard and Mannor where the underlying latent state is fixed, and comes from a finite set of known size. In practice, the number of latent states could be be tuned during offline learning via cross validation. Our algorithms can also work with non-parametric latent models, though parametric ones are more well-studied in literature and have recovery guarantees.

Questions about MovieLens experiment: For each episode/user, actions are movies to recommend, and context is the concatenation of feature vectors for each movie; the context was learned from the training set and known beforehand. The mean reward for recommending movie i to user j is the dot product between movie i and user j's feature vectors, both of which derived from the test set and not known to the learning method. Our hypothesis was that using a Gaussian mixture-model (GMM) over user features fitted on the train set would improve performance by allowing for quickly associating each user's hidden preferences with a cluster in the GMM. We will clarify the ambiguities in our work.

"Exp4 performance seems a bit surprising": Exp4 can solve the latent bandits setting if we interpret the conditional model for each latent state as an expert; however, Exp4 is designed for adversarial rather than stochastic settings, which is why our algorithms greatly outperformed Exp4 using the same GMM model.