ALIGNING MINDS: PITFALLS IN PROXY-BASED MENTAL MODEL ALIGNMENT FOR HUMAN–AI COLLABORATION

Anonymous authors
Paper under double-blind review

ABSTRACT

As AI systems become integral partners in decision-making, mismatches between a user's mental model of the AI and the AI's model of the user can degrade performance, trust, and safety. We introduce Co-Adaptive Mental Model Alignment (CAMMA), a closed-loop framework that jointly infers the user's beliefs about the AI and the AI's estimate of the user's goals in real time, then adapts explanations and solicits feedback to minimize their divergence. In pilot synthetic experiments using an identical-architecture proxy, standard alignment metrics based on Jensen–Shannon divergence saturate near perfection across hyperparameter settings, obscuring meaningful differences. We discuss these negative results as a cautionary example: synthetic proxies and symmetric divergence measures may hide subtle misalignments. We outline future human-subject studies to validate CAMMA and propose refined alignment metrics better suited to real-world collaboration.

1 INTRODUCTION

Effective human–AI collaboration requires mutual understanding: users must form accurate mental models of AI behavior, and AI systems should accurately infer users' goals and expertise. One-way adaptation—either tailoring explanations to a fixed user model or adjusting behavior based on user feedback—addresses only half the problem. We propose CAMMA, a bidirectional loop that alternates between inferring the user's model of the AI and updating the AI's model of the user, guided by an alignment score. Our contributions: (a) definition of the Mutual Model Alignment Score (MMAS) based on Jensen–Shannon divergence; (b) a synthetic pilot study revealing near-saturation of MMAS under identical architectures; (c) a discussion of the limitations of proxy-based alignment metrics and a roadmap for real-user validation.

2 RELATED WORK

User mental models of AI have been studied in HCI and ML. Miller provides a social-science perspective on explanation needs (**?**). Kulesza *et al.* develop explanatory debugging to correct AI misconceptions interactively (**?**). Cooperative IRL frames value alignment as inferring human rewards (**?**). Plan reconciliation methods adjust explanations to satisfy a user's model (**?**). Information-theoretic divergences have been used for alignment assessment (**?**). Trust calibration in automation emphasizes appropriate reliance (**?**). Unlike prior one-way adaptation methods, CAMMA maintains and co-updates both models iteratively.

3 BACKGROUND

For input x, let $P(u \mid x)$ be the AI's model of the user's belief over labels, and $Q(u \mid x)$ the user's belief about the AI. Define

$$\mathrm{JSD}(P, Q) = \tfrac{1}{2}\mathrm{KL}(P\|M) + \tfrac{1}{2}\mathrm{KL}(Q\|M), \quad M = \tfrac{1}{2}(P + Q),$$

and

$$\mathrm{MMAS} = 1 - \mathrm{JSD}(P, Q) \in [0, 1].$$

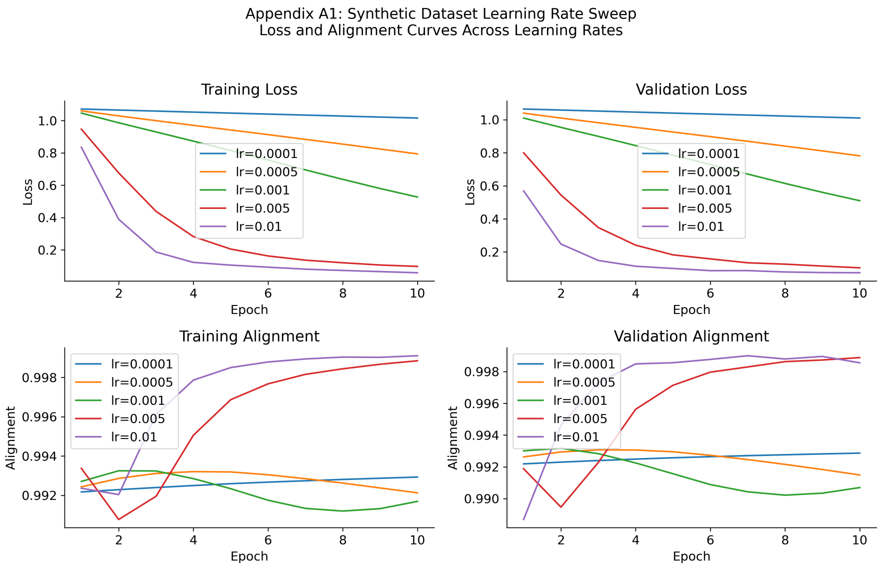

Figure 1: Synthetic dataset learning-rate sweep. Top: training (solid) and validation (dashed) loss vs epoch. Bottom: training (solid) and validation (dashed) MMAS. Higher rates (¿0.005) achieve MMAS¿0.997 by epoch 3, whereas lower rates plateau near 0.992–0.995. A narrow y-range (0.99–1.0) hides these differences.

4 METHOD: CAMMA

CAMMA operates in a continuous loop: (i) mental model inference probes user responses to estimate $Q(u|\, x)$; (ii) user model update applies inverse reinforcement learning (**?**) on observed actions to refine $P(u|\, x)$; (iii) calibration and intervention computes MMAS to trigger adapted explanations (feature-highlighting, natural-language rationales) or feedback requests. Intervention policies trade off alignment gains against interaction cost.

5 EXPERIMENTAL SETUP

We created a synthetic 3-class classification dataset ($N_{\text{train}} = 1000$, $N_{\text{val}} = 200$, $D = 10$). Two 2-layer MLPs (32-unit ReLU hidden layer) were trained with cross-entropy under identical architectures. We swept Adam learning rates in $\{10^{-4}, 5 \times 10^{-4}, 10^{-3}, 5 \times 10^{-3}, 10^{-2}\}$ for 10 epochs and logged losses and MMAS per epoch. Separately, we trained the same MLP on Dbpedia14 with dropout rates $\{0.0, 0.1, 0.3, 0.5\}$ for three epochs, measuring validation loss, MMAS, and Model Alignment Improvement (MAI), defined as MMAS$_t$–MMAS$_{t-1}$.

6 RESULTS

Figure 1 summarizes the synthetic learning-rate sweep. Training/validation loss curves diverge: higher rates converge faster. Alignment (MMAS) curves, however, rapidly rise above 0.99 for all rates and by epoch 3 reach 0.992–0.995 at lower rates versus ¿0.997 at higher rates. The narrow y-axis exaggerates saturation, masking these small but systematic gaps.

Figure 2 shows the Dbpedia14 dropout ablation. Validation loss increases with dropout, while MMAS climbs from 0.990 to 0.996 by epoch 2 and then flattens. MAI exhibits a sharp rise from epoch 1 to 2 (0.94→0.98) but changes little thereafter across dropout rates, indicating marginal sensitivity.

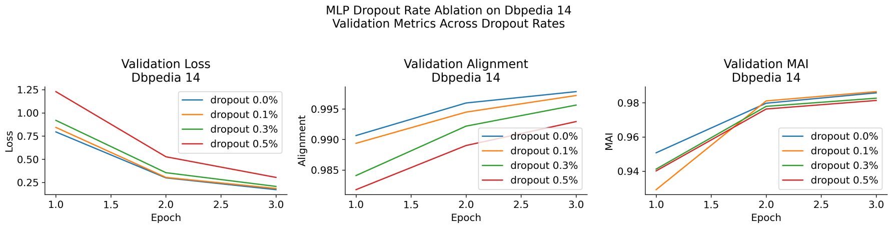

Figure 2: Dbpedia14 MLP dropout ablation. Left: validation loss; middle: MMAS; right: MAI. Dropout markedly affects loss but alignment metrics level off after epoch 2, with only minor spread (dropout 0.0–0.5).

These pilot experiments reveal a pitfall: proxy-based MMAS saturates quickly, obscuring meaningful differences across conditions. Synthetic proxies with identical architectures can yield near-perfect scores, misleading alignment assessments.

7 Conclusion

We introduced CAMMA, a co-adaptive framework for human–AI mental model alignment, and conducted pilot synthetic studies. Our negative findings—rapid saturation of Jensen–Shannon-based MMAS under trivial proxies—highlight the risk of overreliance on symmetric divergence metrics and synthetic benchmarks. Future work must validate CAMMA with human participants on real tasks, develop asymmetric or behavior-driven alignment measures, and design probing policies that uncover subtle model divergences while balancing interaction cost.

Supplementary Material

Implementation Details. Both MLPs used Adam (**?**) with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and the stated learning-rate grid. A single hidden layer of 32 ReLU units was used. Dropout was applied after the hidden layer. Softmax outputs were smoothed with $\epsilon = 10^{-8}$ before divergence computation. All experiments ran on one NVIDIA GPU.

Additional Figures. We omit several auxiliary plots (e.g., optimizer choice, activation distributions) that reinforce MMAS saturation; code and full logs are in the repository. Error-bar and multi-seed analyses are pending real-user validation and thus excluded here to maintain focus on the core pitfall.