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 and the AI's estimate of the user's goals in real time, then adapts explanations and solicits feedback to minimize their divergence. Synthetic pilot studies (see App. A) show standard alignment metrics saturate near perfection, obscuring differences. We then report a dropout ablation on Dbpedia14 revealing minor alignment spreads despite large loss changes. We discuss the limitations of proxy-based metrics, propose refined measures, and plan real-user trials.

1 INTRODUCTION

Effective human–AI collaboration requires mutual understanding: users must form accurate mental models of AI behavior, and AI systems should accurately infer user goals. 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 are: (i) the Mutual Model Alignment Score (MMAS) based on Jensen–Shannon divergence; (ii) synthetic pilots (App. A) revealing MMAS saturation; (iii) a Dbpedia14 dropout ablation (Fig. 3) showing minor alignment shifts despite large loss gaps; (iv) a discussion of proxy-metric pitfalls and a roadmap for human–AI validation.

2 RELATED WORK

User mental models of AI have been studied in HCI and ML. Miller provides social-science insights into 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 assess alignment (**?**). Trust calibration in automation emphasizes appropriate reliance (**?**). Unlike prior one-way adaptation, CAMMA co-updates both models iteratively.

3 METHOD: CAMMA

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

$$\mathrm{JSD}(P,Q) = \tfrac{1}{2}\mathrm{KL}(P\|M)+\tfrac{1}{2}\mathrm{KL}(Q\|M), \quad M = \tfrac{1}{2}(P+Q), \quad \mathrm{MMAS} = 1-\mathrm{JSD}(P,Q) \in [0,1].$$

CAMMA operates in a continuous loop: (i) infer $Q(u \mid x)$ from user responses; (ii) update $P(u \mid x)$ via inverse-IRL (**?**); (iii) compute MMAS to trigger adapted explanations or feedback requests. Intervention policies trade off alignment gains against interaction cost.

4 EXPERIMENTS

We first ran synthetic MLP classification pilots (App. A) to test learning-rate effects on MMAS. Next, we performed a dropout ablation on Dbpedia14 with a 2-layer MLP (32 ReLU units, softmax), Adam (**?**), learning rate 1e-3, dropout $\{0.0,0.1,0.3,0.5\}$, for three epochs. We logged validation loss and MMAS.

5 RESULTS

Synthetic pilots show MMAS ¿0.99 within three epochs for all rates (App. A), with only 0.005 difference—saturation that hides systematic gaps. Figure 3 gives the Dbpedia14 ablation. Validation loss increases with dropout, while MMAS rises to 0.995 by epoch 2 and then plateaus: only ±0.002 spread across dropout rates despite ¿0.5 loss variation. This underscores the limitations of symmetric-divergence metrics. The Model Alignment Improvement (MAI, App. B) similarly shows marginal gains, reinforcing this pitfall.

6 CONCLUSION

We introduced CAMMA, a co-adaptive framework for mental model alignment, and conducted proxy-based pilots. MMAS saturation in both learning-rate and dropout ablations highlights that symmetric divergences on identical-architecture proxies can mask meaningful misalignment. Future work will validate CAMMA with human subjects, develop asymmetric or behavior-driven metrics, and design probing policies to uncover subtle divergences while balancing cost.

A SYNTHETIC LEARNING-RATE SWEEP

B ADDITIONAL MAI CURVE

[b]0.48

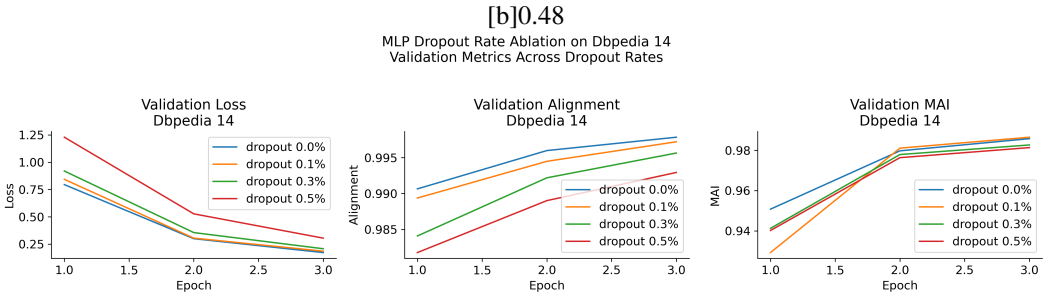

Figure 1: Validation Loss vs. Epoch

[b]0.48

app_alignment_dbpedia14_two_panel.png

Figure 2: Validation MMAS vs. Epoch

Figure 3: Dbpedia14 dropout ablation: (a) higher dropout increases loss; (b) MMAS saturates by epoch 2 with only ±0.002 spread.

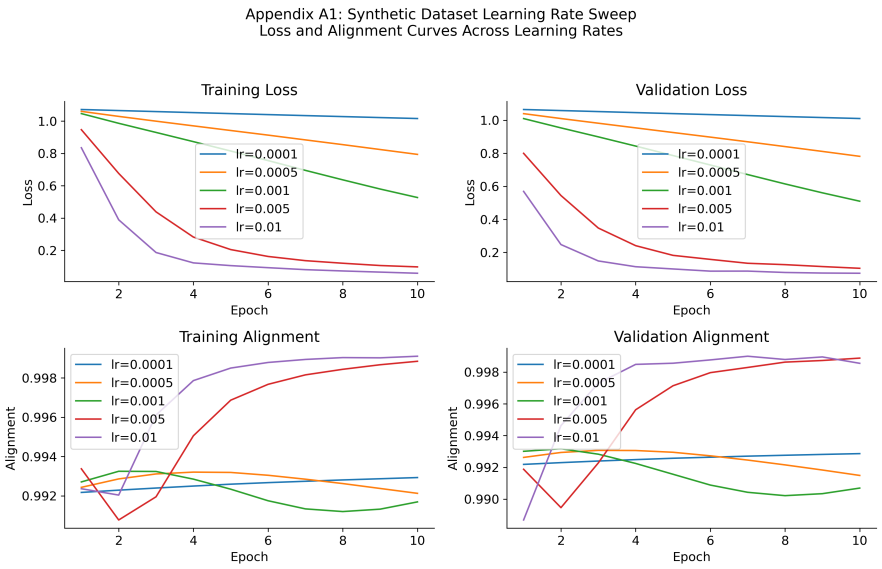

Figure 4: Synthetic 2×2 sweep: top row loss (solid=train, dashed=val); bottom row MMAS. MMAS saturates ¿0.99 by epoch 3.

app_mai_dbpedia14.png

Figure 5: Model Alignment Improvement for Dbpedia14 dropout: sharp rise epoch 1→2, then plateau.