Aligned Machine

Embedding models increasingly inform how people search, recommend, and retrieve information, yet we lack credible ways to test whether these models' similarity judgments reflect how humans understand meaning. Existing datasets such as STS-B (Cer et al., 2017) and SimLex-999 (Hill et al., 2015) rely on static, often expert-labeled pairs with coarse Likert-style annotations. These approaches are limited in scope and vulnerable to inter-annotator inconsistencies.

We introduce Aligned Machine, an interactive, public-facing platform that turns broad participation into a benchmark for evaluating semantic alignment in embedding-based systems. Inspired by participatory studies such as Moral Machine (Awad et al., 2018), the platform invites anyone to compare two pairs of items (text or images) and choose which pair is more similar. This comparative, forced-choice task is cognitively natural for non-experts, creates engaging interactions, and yields clean preference data suitable for model evaluation.

Aligned Machine is designed first and foremost as a vehicle for public engagement and AI literacy. The IRB approved, web-based platform integrates elements of gamification, culminating in an analysis of how a user compared to popular embedding models. By summarizing results in a table and a bar chart, the analysis reveals where user intuitions converge or diverge from different embedding models, encouraging participatory insight into alignment and the difficulty of modeling diverse viewpoints.

Each interaction with the platform also populates a benchmark of human-aligned similarity, which will be released as an open-source dataset and leaderboard. Each record includes the items compared, the human majority choice, individual responses, and model predictions. Because the task uses comparative judgments, standard preference-aggregation methods can recover latent similarity structure from noisy votes, supporting robust evaluation without relying on coarse Likert ratings or static expert labels. This benchmark is intended to support standardized accuracy and ranking across models.

The platform is intentionally designed to grow over time, sustaining public engagement while steadily improving its applied impact. Future integrations will add targeted, domain-specific studies and evaluation tasks that make it possible to detect persistent biases and to measure concept drift as new data and model versions emerge.

By combining accessible participation with rigorous comparative judgments, Aligned Machine provides a scalable pathway to engage the public in the evaluation of AI systems while producing an open, evolving benchmark that the community can use to test, compare, and improve AI models.

References

Awad, E., Dsouza, S., Kim, R. et al. The Moral Machine experiment. Nature 563, 59–64 (2018). https://doi.org/10.1038/s41586-018-0637-6

Cer, D., Diab, M., Agirre, E. et al. SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation. In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), 1–14, Vancouver, Canada. Association for Computational Linguistics (2017).

Hill, F., Reichart, R., Korhonen, A. SimLex-999: Evaluating Semantic Models With (Genuine) Similarity Estimation. Computational Linguistics 41(4), 665–695 (2015).