TRAIL
TRUSTED AI LABS

Summer Workshop 25' London

Aligning Recommendation Explanations to User Preferences Using LLMs Fine-Tuned by Reinforcement Learning with AI Feedback

Project n°6

KING'S College LONDON

KING'S INSTITUTE FOR ARTIFICIAL INTELLIGENCE

Wallonie - Bruxelles International.be

ncp WALLONIE

Wallonie service public SPW

S3

TRAIL
TRUSTED AI LABS

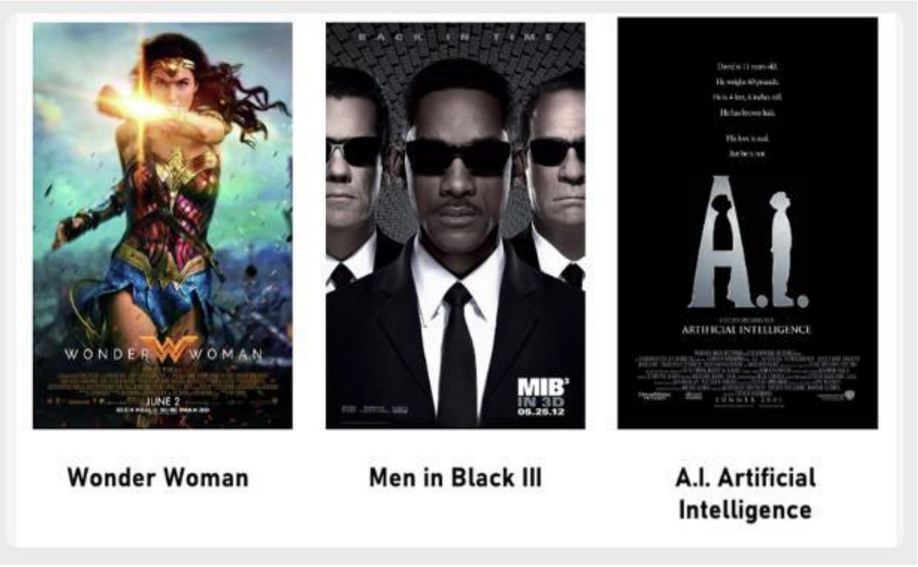
TRAIL

TRUSTED AI LABS

Summer Workshop 25' London

- Yasmine Akaichi, PhD student (UNamur)
- Julien Albert, PhD student (UNamur)
- Martin Balfroid, PhD student (UNamur)
- Lluc Bono Rosselló, PhD student (ULB)
- Lucile Dierckx, PhD student (UCLouvain)
- Sédrick Stassin, PhD (UMONS)
- Vincent Stragier, Research Assistant (UMONS)

Explainable Recommendation



Wonder Woman | Men in Black III | A.I. Artificial Intelligence

Past preferences for Jessica

RECOMMENDATION 3: JESSICA

RECOMMENDATION:

JUSTICE LEAGUE:
Fueled by his restored faith in humanity and inspired by Superman's selfless act, Bruce Wayne enlists the help of his new-found ally, Diana Prince, to face an even greater enemy.

EXPLANATION 1:

We recommended “Justice League” because:

- “Justice League” is from decade Movies of the 2010s like “Men in Black III”

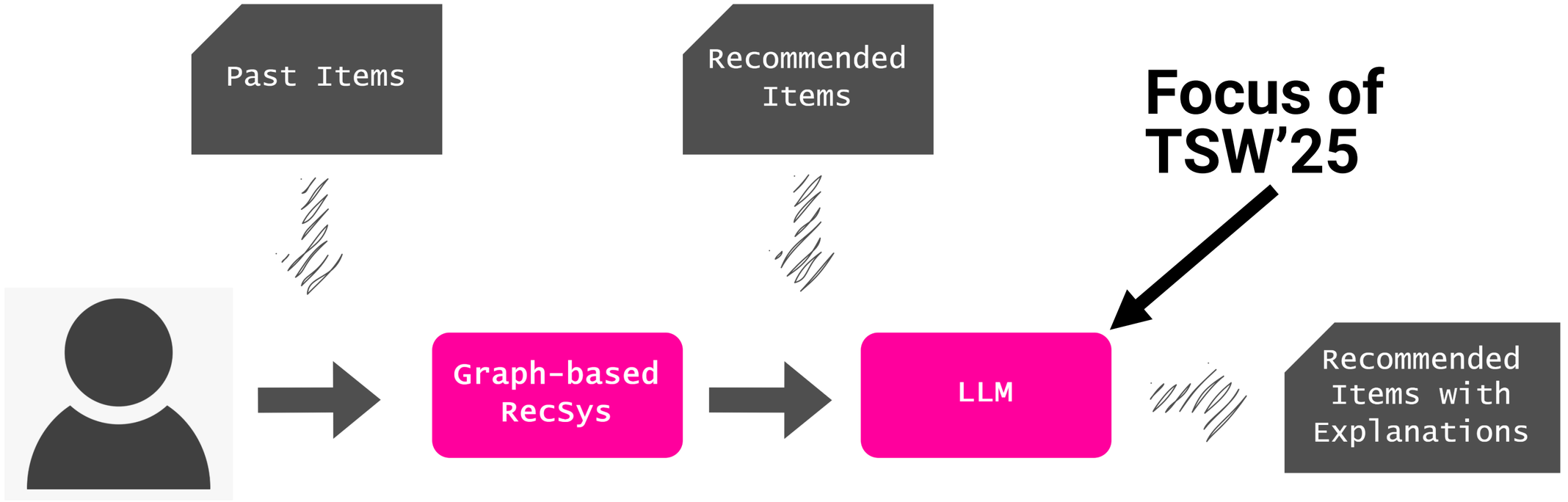
EXPLANATION 2:

We suggested “Justice League” for your viewing pleasure because it is a product of the 2010s decade, similar to “Men in Black III” and we believe it embodies the same level of excitement and adventure that you enjoyed in the latter.

EXPLANATION 3:

Based on the data provided, "Justice League" was likely recommended because it shares several commonalities with other films in the dataset. Specifically, it belongs to the "Decade-2010" category, like "Men in Black III" and "Suicide Squad." Additionally, "Justice League" falls under the "Science Fiction Action Film" genre, which is a subclass of "Science Fiction Film," the same genre as "A.I. Artificial Intelligence" and "Suicide Squad." It is also an "Action Film," like "Men in Black II" and "Wonder Woman." Lastly, "Justice League" is directly related to "Wonder Woman" as it followed that film. All these overlapping attributes make "Justice League" a fitting recommendation given the movies and features specified.

Recommendation Pipeline

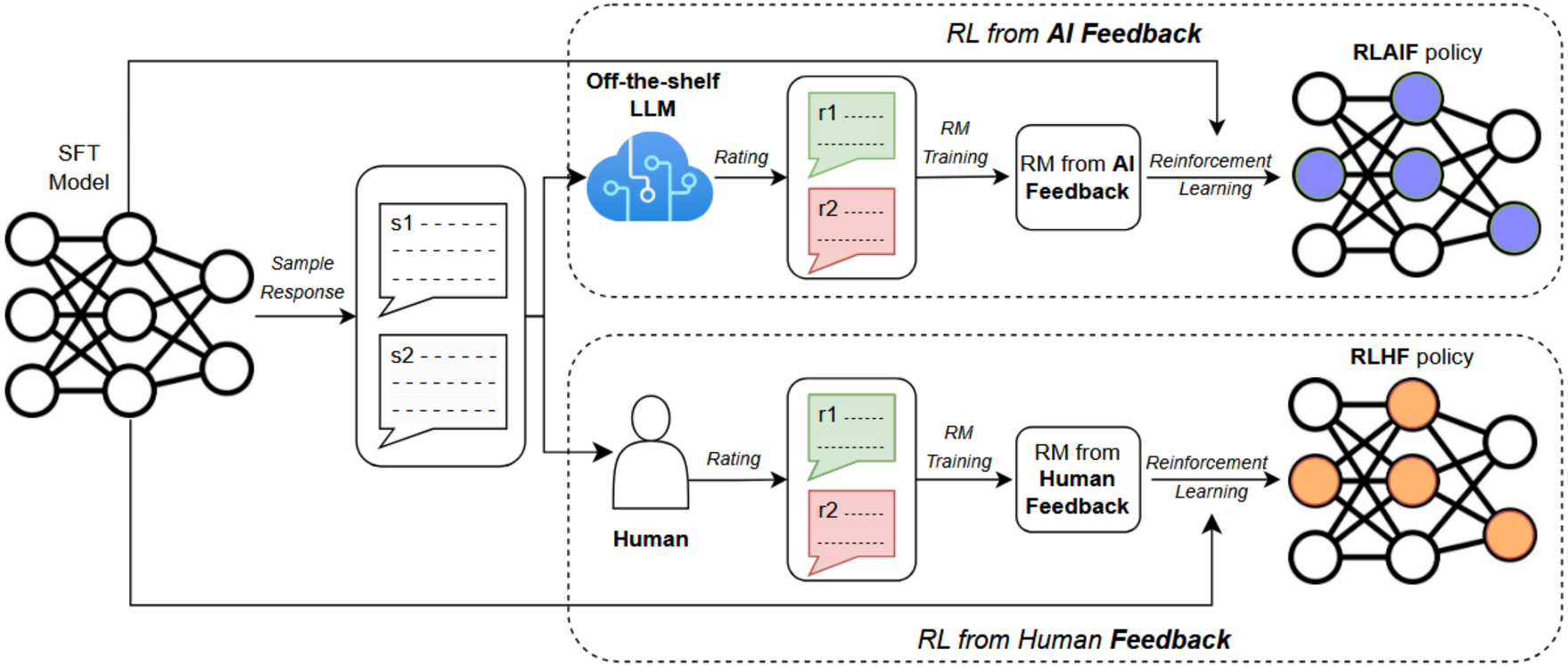


Design & Implementation done during TSW'23 & TSW'24

Reinfor. Learning with AI Feedback

Improve alignment with non-functional requirements

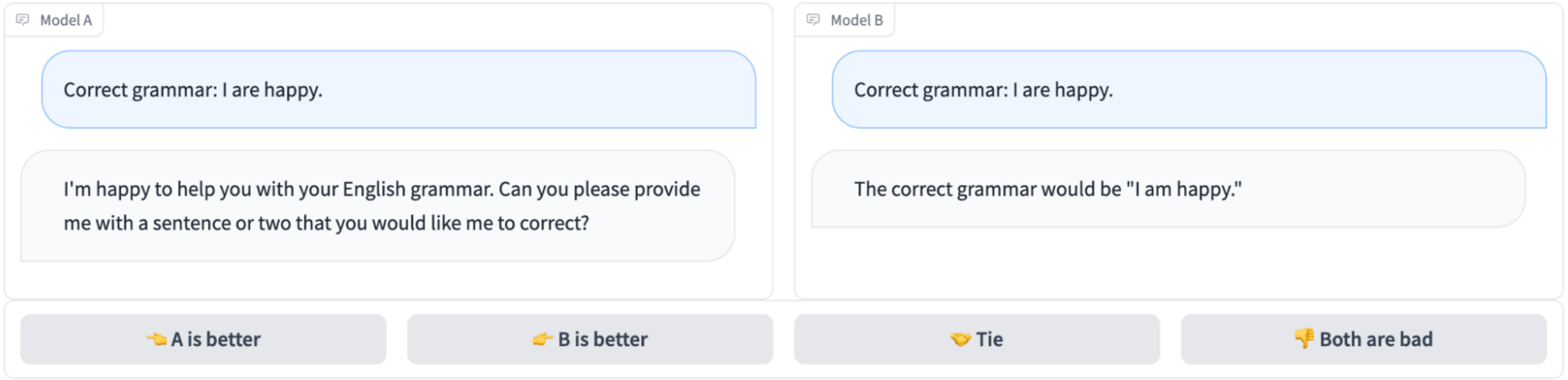
E.g., Satisfaction, Scrutability and Transparency



Lee, Harrison, et al. "RLAIF vs. RLHF: Scaling reinforcement learning from human feedback with ai feedback." arXiv preprint arXiv:2309.00267 (2023). Website

Evaluation of Explainable Recom.

Inspiration from Chatbot Arena



Chiang, Wei-Lin, et al. "Chatbot arena: An open platform for evaluating LLMs by human preference." Forty-first International Conference on Machine Learning. 2024. Website

Work Plan

WP1 – Recommender System [before TSW'25]

- Implementation of the recommendation pipeline
- Integration of graph-based & ILP-based recommendation methods

WP2 – LLM Fine-Tuning with RLAIF

- Design and implementation of the explanation enhancing methods

WP3 – Evaluation of Explainable Recommendations

- Evaluation of the methods using a user-centered approach

WP4 – Communication of the Work

- Communication through a workshop paper

Expertise Sought

Research Axis 1: Recommender Systems

- Graph-based Recommendation
- Inductive Logic Programming
- Explainable Recommendation

Research Axis 2: Large Language Models

- Use of LLMs (prompting strategies, deployment, etc.)
- Reinforcement Learning from Human/AI Feedback

Research Axis 3: User-Based Evaluation

- User Study
- Online Evaluation

TRAIL
TRUSTED AI LABS

Summer Workshop 25’ London

Thank you for your attention !