Institute of
Artificial Intelligence

Leibniz
Universität
Hannover

Developing Interpretable Style Vectors to Steer Large Language Models towards Group-Specific Explanation Generation

Janek Prange

08.04.2025

Table of Contents

Table of Contents

Interpretable Style Embeddings

Group-Specific Explanations

Table of Contents

Group-Specific Texts

Table of Contents

Table of Contents

Table of Contents

Table of Contents

Thank you for your attention

References I