



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



Table of Contents

Motivation

Data Collection

Creation of the Style Vector Attributes

- Style Sentence Generation

- Clustering and Cluster Selection

Custom Models

- SFAM

- LISA

- Embedding Model

Steering Text Generation

- Steering with Prompt Engineering

- Activation Steering

Conclusion

Table of Contents

Motivation

Data Collection

Creation of the Style Vector Attributes

Style Sentence Generation

Clustering and Cluster Selection

Custom Models

SFAM

LISA

Embedding Model

Steering Text Generation

Steering with Prompt Engineering

Activation Steering

Conclusion

Interpretable Style Embeddings

Group-Specific Explanations

Table of Contents

Motivation

Data Collection

Creation of the Style Vector Attributes

Style Sentence Generation

Clustering and Cluster Selection

Custom Models

SFAM

LISA

Embedding Model

Steering Text Generation

Steering with Prompt Engineering

Activation Steering

Conclusion

Group-Specific Texts

Table of Contents

Motivation

Data Collection

Creation of the Style Vector Attributes

Style Sentence Generation

Clustering and Cluster Selection

Custom Models

SFAM

LISA

Embedding Model

Steering Text Generation

Steering with Prompt Engineering

Activation Steering

Conclusion

Table of Contents

Motivation

Data Collection

Creation of the Style Vector Attributes

Style Sentence Generation

Clustering and Cluster Selection

Custom Models

SFAM

LISA

Embedding Model

Steering Text Generation

Steering with Prompt Engineering

Activation Steering

Conclusion

Table of Contents

Motivation

Data Collection

Creation of the Style Vector Attributes

Style Sentence Generation

Clustering and Cluster Selection

Custom Models

SFAM

LISA

Embedding Model

Steering Text Generation

Steering with Prompt Engineering

Activation Steering

Conclusion

Table of Contents

Motivation

Data Collection

Creation of the Style Vector Attributes

Style Sentence Generation

Clustering and Cluster Selection

Custom Models

SFAM

LISA

Embedding Model

Steering Text Generation

Steering with Prompt Engineering

Activation Steering

Conclusion

Thank you for your attention

References I