Appendix

A Evaluation Prompts

In this investigation, we compared the system's performance on *factual* and *subjective* on prompts. Comprehensive lists of these prompts are provided in Sec. A.1 and Sec. A.2, respectively.

A.1 Factual Prompts

There were 50 factual prompts used in this study, which are referred to as **F01** to **F50**:

[F01] How many bones are there in the human body?

[F02] How many chambers are there in the human heart?

[F03] How many elements are there in the periodic table?

[F04] How many planets are there in our solar system?

[F05] How many players are there in a baseball team?

[F06] How many players are there in a volleyball team?

[F07] How many symphonies did Ludwig van Beethoven compose?

[F08] In which year did World War II end?

[F09] In which year did the Berlin Wall fall?

[F10] In which year did the first moon landing occur?

[F11] What is the boiling point of water in Fahrenheit?

[F12] What is the capital city of France?

[F13] What is the chemical formula for methane?

[F14] What is the chemical formula for table salt?

[F15] What is the chemical formula for water?

[F16] What is the chemical symbol for gold?

[F17] What is the chemical symbol for sodium?

[F18] What is the deepest point in the Earth's oceans?

[F19] What is the formula for calculating density?

[F20] What is the formula for calculating the area of a circle?

[F21] What is the formula for calculating the area of a triangle?

[F22] What is the formula for calculating the volume of a cylinder?

[F23] What is the formula for converting Celsius to Fahrenheit?

[F24] What is the freezing point of water in Kelvin?

[F25] What is the largest country in the world by land area?

[F26] What is the largest internal organ in the human body?

[F27] What is the largest ocean in the world?

[F28] What is the largest organ in the human body?

[F29] What is the speed of light in a vacuum?

[F30] What is the symbol for the chemical element iron?

[F31] What is the tallest building in the world?

[F32] What is the tallest mountain in the world?

[F33] What is the world's longest river?

[F34] Which country is famous for the Taj Mahal?

[F35] Which country is known as the Land of the Rising Sun?

[F36] Which gas is known as laughing gas?

[F37] Which gas makes up the majority of Earth's atmosphere?

[F38] Who developed the theory of evolution by natural selection?

[F39] Who discovered penicillin?

[F40] Who discovered the theory of general relativity?

[F41] Who is considered the father of modern physics?

[F42] Who is credited with inventing the telephone?

[F43] Who is the author of the play "Romeo and Juliet"?

[F44] Who is the current President of the United States?

[F45] Who painted "The Starry Night"?

[F46] Who painted the "Last Supper"?

[F47] Who painted the Mona Lisa?

[F48] Who wrote the novel "Pride and Prejudice"?

[F49] Who wrote the novel "To Kill a Mockingbird"?

[F50] Who wrote the play "Hamlet"?

A.2 Subjective Prompts

The 49 applied factual prompts are referred to as **S01** to **S49**:

[S01] Announce the weather forecast for the upcoming weekend.

[S02] Ask your hairdresser for an appointment next week to have your hair dyed.

[S03] Comment on a critical review of a customer of your business.

[S04] Compare the color blue and green.

[S05] Compare the cultural value of theaters and cinemas.

[S06] Compare the qualities of coffee and tea.

[S07] Compare the relaxation based on vacation and continuous sport.

[S08] Compare the taste of a strawberry smoothie to that of a vanilla one.

[S09] Compose a few lines of lyrics talking about society.

[S10] Describe a fictional character.

[S11] Describe a meal or dish that holds sentimental value to you and why.

[S12] Describe a person who has had an impact on your life and why.

[S13] Describe a piece of artwork.

[S14] Describe an incident that could lead to an airplane crash in mid-flight.

[S15] Discuss the impact of social media on interpersonal relationships.

[S16] How can I learn about Machine Learning most efficiently?

[S17] How do caterpillars turn into butterflies?

[S18] How do you approach decision-making when faced with multiple options?

[S19] How do you define art?

[S20] How do you define happiness?

[S21] How do you define sadness?

[S22] How do you feel about the death penalty?

[S23] How do you prioritize your tasks and responsibilities in your daily life?

[S24] How do you stay motivated and focused on long-term goals?

[S25] How would you handle a disagreement with a close friend?

[S26] How would you respond to a rude customer in a customer service role?

[S27] If a roommate consistently borrows your belongings without asking, how would you handle it?

[S28] Order a vegan dish from the menu of a steak house.

[S29] Review the pair of headphones that I bought online last week.

[S30] Share a personal anecdote about a vacation you had.

[S31] Share a personal experience of overcoming a fear or facing a challenge.

[S32] Share a personal experience that challenged your beliefs and changed your perspective.

[S33] Share a quote or mantra that inspires you and explain why it resonates with you.

[S34] Tell me a joke.

[S35] Tell me about the concert in America last year.

[S36] What approach would you take when faced with a challenging work colleague?

[S37] What approaches do you take to foster creativity and innovation in your work?

[S38] What are the steps to bake a perfect pizza from scratch?

[S39] What did a day in a typical family in the year 1980 look like?

[S40] What do German bread rolls taste like?

[S41] What does a fulfilling and meaningful life mean to you?

[S42] What is your opinion on the role of government in society?

[S43] What is your perspective on the significance of cultural heritage?

[S44] What is your preferred method of learning new skills or acquiring knowledge?

[S45] What is your stance on the role of machine learning in education?

[S46] What strategies do you find most effective for managing stress?

[S47] Would you like to see a movie about yourself?

[S48] Write a recipe to make chocolate chip muffins.

[S49] Write an introduction about yourself for a CV.

The full list of produced answers to the prompts in Sec. A.1 and Sec. A.2 can be found at [redacted for peer-review].

B Effect of the parameter λ

The example below shows an example of the effect of the weighting parameter lambda for style vectors (Eq. 3) on steering a response toward negative sentiment. Larger values generally lead to increased negativity. However, if the values are too large, the model's output is not adequate anymore (oversteering).

P: *Review the pair of headphones that I bought online last week.*

$\lambda = 0$**:** *I recently bought a pair of headphones online and I am very pleased with them. The sound quality is excellent and they are very comfortable to wear. The design is also stylish and the adjustable headband allows for a custom fit. I would highly recommend these headphones to anyone looking for a good pair of audio headphones.*

$\lambda = 0.6$**:** *I recently bought a pair of headphones online and I'm very satisfied with them. The sound quality is great and they are very comfortable to wear. The design is also very stylish and I love the adjustable fit. Highly recommended!*

$\lambda = 1.2$**:** *I love the pair of headphones you got! The sound quality is great and the comfortable design is perfect for all day listening. High recommend!*

$\lambda = 1.9$**:** *<s> great great great great great great ...*

C Further results from the probing study

Analogously to the analysis of the Yelp dataset in Sec. 4.3, we performed the same experiment with the Shakespeare and the GoEmotions datasets.

Shakespeare The capabilities of the trained steering vectors $\mathbf{z}_{\mathbf{x}}^{(\mathbf{i})}$ and activations $\mathbf{a}^{(\mathbf{i})}(\mathbf{x})$ at layer i to encode style in the Shakespeare dataset are presented in Fig. 6. In contrast to the Yelp review dataset, we want to differentiate between modern and original Shakespearean phrases. This task differs from the other two datasets in that we do not change emotion or sentiment but a whole writing style. The Shakespeare classifier on the trained steering vectors reaches a maximal AUC value of 0.8, while their corresponding activation vectors reach an AUC value of 0.96. Again, the layers $i \in \{18, 19, 20\}$ had high AUC values. This supports our initial findings on the Yelp review dataset. As can be seen by comparing the AUC values for the activation vectors from Shakespeare (max. AUC = 0.96/ Fig. 6c) with Yelp in the same setting (max. AUC = 0.99/ Fig. 6c), the style difference between original and modern Shakespeare is harder to distinguish, than the sentiment in the Yelp reviews.

GoEmotions For this dataset, the ROC plots need to be compared per layer because there are six instead of not two classes. The results for layer 19 draw a slightly different picture (Fig. 8) than for Yelp and Shakespeare. Probing the activations of all samples still results in the best micro-average AUC of 0.90. However, in the fair comparison (activations for the 89 samples for which trained steering vectors exist), they have a micro-average AUC of 0.74, while the corresponding trained vectors reach an AUC of 0.82. Nevertheless, this can also result from the small number of trained steering vectors found. The same result can be seen for layers 18 (Fig. 7) and 20 (Fig. 9). We need to investigate this finding in future studies to rule out a statistical anomaly as the cause for this. Still, the layers $i \in \{18, 19, 20\}$ have high micro-average AUC values of around 0.91 for all activations and 0.81 for the trained steering vectors.

Classifier training During our experiments, we tried training the regression model in three different settings: Predicting the class using only a single layer, using three subsequent layers, and training on all layers together. The difference between the resulting classifications is minimal, albeit performance slightly increases when using more layers. For ease of presentation and readability of the plots, we decided to only include single-layer classifiers.