Adaptive Prompting: Ad-hoc Prompt Composition for Social Bias Detection

Maximilian Spliethöver [1], Tim Knebler [1], Fabian Fumagalli [2], Maximilian Muschalik [3], Barbara Hammer [2], Eyke Hüllermeier [3], Henning Wachsmuth [1]
[1]Leibniz University Hannover, Institute of Artificial Intelligence
[2]Bielefeld University, CITEC
[3]LMU Munich, MCML
m.spliethoever@ai.uni-hannover.de

Abstract

Recent advances on instruction fine-tuning have led to the development of various prompting techniques for large language models, such as explicit reasoning steps. However, the success of techniques depends on various parameters, such as the task, language model, and context provided. Finding an effective prompt is, therefore, often a trial-and-error process. Most existing approaches to automatic prompting aim to optimize individual techniques instead of compositions of techniques and their dependence on the input. To fill this gap, we propose an *adaptive prompting* approach that predicts the optimal prompt composition ad-hoc for a given input. We apply our approach to social bias detection, a highly context-dependent task that requires semantic understanding. We evaluate it with three large language models on three datasets, comparing compositions to individual techniques and other baselines. The results underline the importance of finding an effective prompt composition. Our approach robustly ensures high detection performance, and is best in several settings. Moreover, first experiments on other tasks support its generalizability.

1 Introduction

The development of instruction-tuned large language models (LLMs) has led to an increased interest in prompting techniques to augment inputs (Tian et al., 2024). Whereas prompts for earlier generative language models consisted only of the input text and special tokens (Radford et al., 2018; Raffel et al., 2020), they may now encompass complex and explicit task instructions with ancillary information. Among others, successful prompting techniques include personas (Liu et al., 2024a), in-context demonstrations (Liu et al., 2023; Dong et al., 2024), and reasoning steps (Wei et al., 2022).

The success and applicability of prompting techniques does, however, depend on several parameters, including the target task, the language model and its size, and the context provided (Brown et al., 2020; Schick et al., 2021; Mosbach et al., 2023; Dong et al., 2024; Arroyo et al., 2024). Finding an effective prompt is, therefore, often still a time-consuming process that needs re-evaluation should any of the parameters change. Recent automated prompting methods either address the lexical aspect by finding the best formulation (Honovich et al., 2023) or manipulate the latent space (Liu et al., 2022b). Most automatic methods, however, focus on optimizing a single technique and do not consider a composition of techniques that could take advantage of their individual strengths.

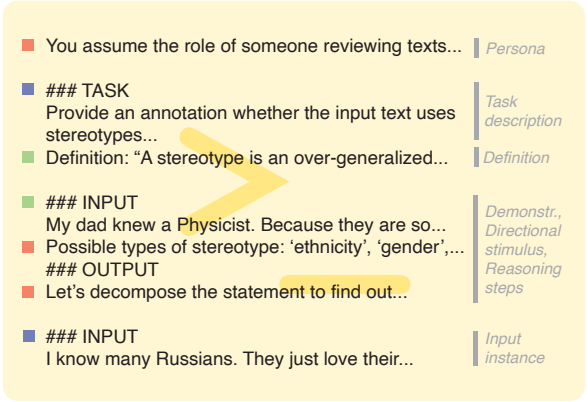

Figure 1: Exemplary excerpt of a prompt composition for social bias detection. While certain techniques might benefit detection performance (say, those with green squares), others might not (red squares). Some parts always need to be present (blue squares). A full prompt example is shown in Figure 10.

The example in Figure 1 visualizes a prompt composition of several techniques. While a *definition* provides a theoretical background, *in-context demonstrations* clarify its application. As recent works indicate that LLMs cannot attend equally to all available information (Liu et al., 2024b; Plepi et al., 2024), simply using more techniques may add noise and reduce performance. Finding an optimal prompt composition is, therefore, desirable.

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 2421–2449
April 29 - May 4, 2025

To this end, we propose an *adaptive prompting* approach to predict the optimal composition of discrete prompt techniques *ad-hoc*, i.e., for each input instance. First, an encoder model learns to predict *optimal compositions* based on a pool of individual prompting techniques. Consecutively, the approach predicts the best composition for each instance and prompts the LLM accordingly.

We evaluate adaptive prompting for five techniques and their compositions for the task of social bias detection. The task requires semantic understanding and world knowledge (Zhou et al., 2023a), likely benefitting from using multiple prompting techniques, making it a good candidate to evaluate prompt compositions. To better understand the importance of each technique and their second-order interactions, we further conduct a Shapley interaction analysis (Fumagalli et al., 2023).

We test our adaptive prompting approach on three social bias datasets for three open-weight LLMs, namely Mistral (7B) (Jiang et al., 2023), Command-R (35B) (CohereForAI, 2024), and Llama 3 (70B) (Dubey et al., 2024). We compare to baselines within and across datasets, including a fine-tuned model and a composition ensemble. The results suggest that our approach robustly ensures high classification performance; in many cases, it even outperforms all baselines and fixed compositions. Moreover, follow-up experiments on three other NLP tasks stress the generalizability of our approach beyond social bias.

To summarize, our main contributions are:

- We present a novel discrete prompt optimization approach to predict the optimal prompt composition for a given model and input. It improves over several baselines and individual prompting techniques on selected datasets.
- We evaluate the utility of generative LLMs and prompting for social bias detection for three state-of-the-art LLMs on three datasets and with five prompting techniques.
- We provide insights into the performance and interaction of prompting techniques, finding that well-performing techniques can also interact negatively when used with others.[1]

2 Related work

Discrete prompts for LLMs have been evaluated for various tasks and applications. For example, Zamfirescu-Pereira et al. (2023) and Arroyo et al. (2024) investigate how sub-optimal prompts affect outputs and Hida et al. (2024) study how different prompts influence social bias exhibited by LLMs.

Several techniques have been proposed to optimize discrete prompts. Popular techniques include personas for perspective taking (Sheng et al., 2021; Xu et al., 2023; Liu et al., 2024a), in-context demonstrations to provide application examples (Dong et al., 2024), and reasoning steps that divide the tasks into sub-tasks (Wei et al., 2022). In this work, we do not consider single prompting techniques but rather evaluate prompt compositions to take advantage of several techniques. Some related works also explore the effect of combining techniques (Stahl et al., 2024), aiming to find the best composition on average. In contrast, we aim to find the optimal composition on each text.

Among existing approaches to automatic prompt optimization, continuous prompt optimization learns to adjust the latent space (Li and Liang, 2021; Liu et al., 2022b), whereas several studies generate discrete optimized instructions from task examples, to easier adapt to unseen data (Zhou et al., 2023b; Honovich et al., 2023; Ha et al., 2023) or new LLMs (Memon et al., 2024). Other works focus on iteratively optimizing prompts (Zhang et al., 2022; Shum et al., 2023; Tian et al., 2024) or predicting the suitability of prompts (Yang et al., 2024). Instead of optimizing the latent space or single techniques, we propose to automatically find optimal prompt compositions for unseen inputs.

Related to the idea of prompt compositions, Khattab et al. (2023) propose a framework to optimize the use of multiple techniques by training a parameterized model that automatically optimizes the prompt. We do not optimize techniques but learn to predict the optimal composition of techniques for a given setting. We further analyze the importance of each technique using Shapley values.

While several other studies aim to detect social bias in text corpora (Spliethöver and Wachsmuth, 2020; Asr et al., 2021; Toro Isaza et al., 2023; Derner et al., 2024), we aim to identify bias in single text instances, similar to Schick et al. (2021); Spliethöver et al. (2024); Powers et al. (2024), by optimizing prompt compositions.

Detecting social bias reliably requires understanding of social language and pragmatics to interpret the implications of text (Choi et al., 2023). Hovy and Yang (2021) identify seven factors of language (e.g., receiver information) to successfully

[1]Code at: https://github.com/webis-de/NAACL-25.

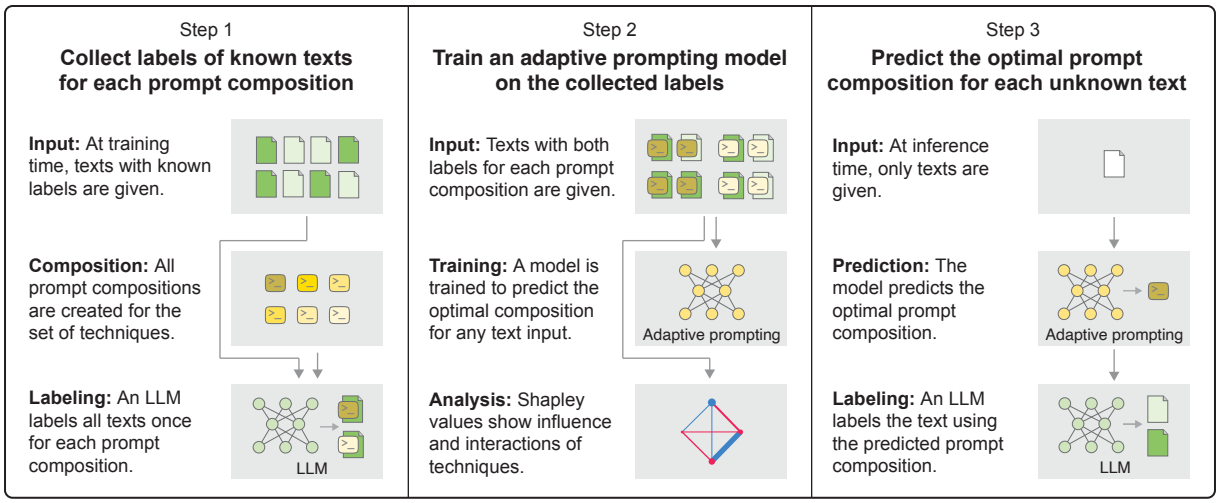

Figure 2: The three steps of our adaptive prompting approach: (1) Bias labels are collected for all considered prompt compositions. (2) A model is trained on the collected labels to predict the optimal composition for any given text. (3) Given an unknown text, the model is applied to predict and use the optimal prompt composition for that text.

model social aspects. Choi et al. (2023) and Zhou et al. (2023a) design social language benchmarks and find that LLMs are still limited in this regard. In this work, prompt compositions enable the inclusion of social aspects (e.g., receiver information with persona prompts) and can, therefore, provide helpful context for social language tasks.

3 Approach

We evaluate instruction-tuned LLMs to identify social bias using compositions of prompting techniques. Specifically, we propose a three-step approach to predict the optimal composition ad-hoc per input text. The process is illustrated in Figure 2.

The first step is to create prompt compositions that represent order-sensitive combinations of one or more prompting techniques. We prompt an LLM with each of these compositions to collect social bias classification labels. In the second step, we use the collected labels to train an adaptive prompting model that finds the optimal prompt composition for a given text instance. Aside from the main task, we also conduct a Shapley value analysis to determine the importance of prompting techniques. We then apply adaptive prompting to unknown texts to find the optimal prompt composition for each.

3.1 Composition and Label Collection

We start by collecting bias label predictions for all possible prompt compositions, emanating from our base composition, a set of prompting techniques, and constraints for their ordering and compatibility.

Base Composition As the minimal prompt to solve the task, we consider the *task description* and *input instance* to be always present (cf. Figure 1).

Ordering and Compatibility Constraints Not all orderings and combinations of prompting techniques create meaningful prompts. For example, if both a *definition* (e.g., of social bias) and *in-context demonstrations* (e.g., a biased text) are present, demonstrations make sense after the definition only. To ensure meaningful prompts, we pre-define a general ordering. Further, variants of the same technique (e.g., demonstrations sampled randomly or by similarity) should not appear together in the same composition; they are mutually exclusive.

Prompting Techniques Let a set of $n \geq 1$ prompting techniques, $T = \{t_1, \ldots, t_n\}$, be given with fixed order t_i before t_j, if $i < j$. We distinguish techniques with one variant, T_1, and with multiple variants, T_2, that is, $T = T_1 \sqcup T_2$, where each $t \in T_2$ has $|t|$ variants. Concretely, each technique in T_1 may or may not be used, and any or none of the techniques in T_2 may be used. Let the set of distinct compositions be denoted as C. Then, the number of compositions is

$$|C| = 2^{|T_1|} \cdot \prod_{t \in T_2} (|t| + 1). \quad (1)$$

3.2 Composition Prediction and Selection

As visualized in the second step of Figure 2, we use the results of the label collection described above to train a prompt composition *prediction model*. The model is then used for *adaptive prompting* (Step 3).

Prediction Model The model is an encoder model with a regression head that is fine-tuned on the collected social bias labels to predict the *optimal* prompt composition c_o from the set of possible compositions C. Here, *optimal* refers to the composition with the highest predicted likelihood to generate a correct bias label for an input. We formulate the prediction of the optimal prompt composition as a regression problem using a sigmoid output layer, followed by binary cross-entropy loss (Ridnik et al., 2021; Grivas et al., 2024). The model's output is a $|C|$-dimensional vector, $\hat{y}$, in which each value represents an independent likelihood estimation for one of the $|C|$ compositions to be optimal.[2]

Adaptive Prompting Given an unknown text, the composition with the highest likelihood is assumed to be an optimal prompt composition c_o, where $o := \arg\max(\hat{y})$. Thereby, we adaptively select a prompt depending on the text at hand.

3.3 Shapley-Based Composition Analysis

Analyzing the impact of individual prompting techniques and their interactions is crucial to evaluate outputs of the adaptive prompting model. To gain these insights, we rely on Shapley values (Shapley, 1953), modeling the predictive performance across all possible compositions as a cooperative game:

Prompt Composition Game Given the set of techniques T, let $\nu : 2^T \rightarrow \mathbb{R}$ be the performance of the techniques $T_c \subseteq T$ of a composition c:

$$\nu(c) := \lambda(y, \hat{y}_c) \tag{2}$$

Here, λ is a performance metric, y are the ground-truth bias labels, and $\hat{y}_c$ the predictions of c. For techniques in T_2, a specific choice must be fixed.

We compute one Shapley value (SV) for each prompting technique $t \in T$, which provides contribution values $\phi(t)$. The SV quantifies the impact of a prompting technique across all possible compositions. Beyond individual contributions, we compute pairwise Shapley interactions (Lundberg et al., 2020) that additionally assign contributions to all pairs of techniques. Shapley interactions (SIs) reveal *synergies* and *redundancies* among prompting techniques, capturing the behavior of the game with greater fidelity (Tsai et al., 2023; Fumagalli et al., 2024b). Akin to Shapley-based feature or data selection (Rozemberczki et al., 2022), we select optimal compositions based on SVs and SIs.

[2]We refrain from a multi-class setup, where the optimal composition is determined over a probability distribution spanning all compositions to avoid a few dominant compositions from possibly being preferred over others consistently.

4 Experiments

In this section, we detail the adaptive prompting experiments that we carried out for the task of social bias detection on three datasets with three state-of-the-art instruction-tuned LLMs. We selected five common prompting techniques that fit the task, which we combine to create prompt compositions. We evaluate our approach against own and related-work baselines, both within and across datasets.

4.1 Task

Social bias detection describes the task of identifying texts that induce bias against a particular social group, through offensive language, stereotypes (Nadeem et al., 2021), power dynamics (Sap et al., 2020; Zhou et al., 2023a), or similar. More context can support the bias detection, making it a well-suited task to evaluate prompt compositions.

In particular, the task requires knowledge about the state of the world to understand implicit biases and dynamics (Hovy and Yang, 2021; Zhou et al., 2023a). Prompting techniques such as in-context demonstrations can show how to apply knowledge acquired during pre-training to identify implicit biases. Furthermore, what is considered a biased text can vary, e.g., based on the target application. Including definitions of bias for and deducting reasoning steps could help to clarify the context and make the predictions more reliable. Lastly, whether a text is considered biased partly depends on the receiver. Instructing the LLM to assume a specific persona can clarify the evaluating perspective.

4.2 Data

To cover multiple aspects of social bias, we evaluate prompt compositions and their predictability on three datasets covering diverse intentions and target applications. In the following, we briefly describe each dataset and how we derive binary bias labels (preprocessing and prompt details in Appendix B).

StereoSet The StereoSet corpus (Nadeem et al., 2021) serves the evaluation of stereotypes in generative models. Here, stereotypes are "over-generalized [beliefs] about a particular group of people" (Nadeem et al., 2021). The data consists of scenarios and target groups that can be combined to create *stereotypical* (biased), *anti-stereotypical*, and *meaningless* (both not biased) texts.

SBIC The Social Bias Inference Corpus (SBIC) Sap et al. (2020) is intended to model multiple aspects of social bias explicitly. We use the *implicit bias* aspect as the target label, which indicates text that are offensive towards a specific social group.

CobraFrames The CobraFrames corpus (Zhou et al., 2023a) captures the social and situational context of biased statements. Among others, it captures bias as implicit power dynamics or stereotypes between the speaker and the listener. We follow Zhou et al. (2023a) in converting the *offensiveness* label into binary bias representations.

4.3 Prompting Techniques

We select five common prompting techniques to investigate potential benefits of prompt compositions over single techniques for social bias detection. We focus on discrete prompting techniques as opposed to continuous methods, such as prefix-tuning (Li and Liang, 2021) or p-tuning (Liu et al., 2022b). Notice, though, that our adaptive prompting is applicable to arbitrary techniques in principle.

Since we evaluate prompt compositions across three datasets with different structures and definitions of social bias, specific content parts of the prompts are adjusted to better align with the scenario of each dataset. In the following, we briefly describe each technique, including dataset alignments and mutations. See Appendix B for details on their lexical representations and a full example.

Personas This technique aims to instruct the model to follow consecutive instructions from the perspective of a specific persona (Thoppilan et al., 2022; Deshpande et al., 2023). Among other use cases, personas are used to build translation systems (He, 2024) and dialogue agents (Thoppilan et al., 2022; Xu et al., 2023), or to investigate biases in pre-trained LLMs (Beck et al., 2024).

For social bias detection, a persona can help clarify the perspective from which a given text is being judged (Giorgi et al., 2024). Since the intentions and goals vary across datasets, we expect that a persona prompt can clarify the setting of the task. In our experiments, we aim to formulate the persona description as close as possible to the scenario envisioned for each dataset. Similar to Xu et al. (2023), we seek to minimize positionality bias.

Definitions Including a definition of social bias can be seen as an extension of the task description to further specify its subject. This is supposed to make the interpretation of social bias in the respective dataset explicit, which is otherwise learned implicitly only in a supervised learning setup. Recent research has found that such definitions can increase the prediction performance in low-resource settings (Elsner and Needle, 2023).

If the authors provide an explicit definition of bias, we reuse it. Otherwise, we manually derive a definition from available information.

In-context Demonstrations Known also as few-shot examples, this technique is a form of in-context learning that "allows language models to learn tasks given only a few examples" (Dong et al., 2024). In-context demonstrations are provided during inference as part of the prompt, unlike traditional fine-tuning where model parameters are optimized in a supervised learning phase (Mosbach et al., 2023). In-context demonstrations have been shown to improve results on various target tasks (Zamfirescu-Pereira et al., 2023; Dong et al., 2024).

While seemingly simple, this technique entails several points of variation, including the *number* and *selection* of examples (Liu et al., 2022a; Zhang et al., 2022; Levy et al., 2023; Bertsch et al., 2024; Dong et al., 2024) and their *ordering* (Lu et al., 2022; Shum et al., 2023; Liu et al., 2024b). To keep the experiments conceivable, we use one demonstration per bias type, adjust the number of similar demonstrations accordingly, and do not evaluate the ordering, but we focus on three common variations to select demonstrations:

- *Random.* We pseudo-randomly select training instances, which are used as demonstrations for all instances in the test split.
- *Similarity.* For each instance in the test split, we select the most similar instances from the training split as demonstrations, using SBERT (Reimers and Gurevych, 2019).
- *Category.* We select instances that cover all bias types covered in a dataset, as diversifying demonstrations has been shown to aid prediction (Levy et al., 2023; Zhang et al., 2022).

Directional Stimulus Directional stimuli (Li et al., 2023) describe the technique to include instance-specific hints and are meant to guide the LLM. We include a list of dataset-specific bias types that could be present in the text instance.

Reasoning Step Instructions Initially intended for "arithmetic, commonsense, and symbolic rea-

soning tasks" (Wei et al., 2022), the main idea of this technique is to decompose the given task into smaller, more approachable sub-tasks (Dong et al., 2024). Reasoning step instructions have been applied to various tasks and can lead to improvements in prediction performance (Wei et al., 2022).

As the detection of social bias in texts can often naturally be decomposed into step-wise questions (Sap et al., 2020; Zhou et al., 2023a), we include reasoning steps as an additional technique. We evaluate both zero-shot and few-shot settings. To do so, we follow Press et al. (2023) and Zhou et al. (2022) in formulating the reasoning steps as task- and data-specific sub-questions covering the aspects of social bias in the respective dataset before prediction. To ensure that all predefined reasoning steps are followed as intended, we separate the reasoning steps into multiple consecutive inference steps (Dong et al., 2024), implemented as a practical chain-of-prompts pipeline (Zhou et al., 2022).

4.4 Models

We realize our *adaptive prompting* approach for three different *instruction-tuned LLMs* as follows.

Adaptive Prompting Given the five prompting techniques, we fine-tune a DeBERTA-v3-large encoder model (He et al., 2023) to predict the optimal composition ad-hoc, i.e., for a given input, as detailed in Section 3. Since the prompting techniques include three variants of in-context demonstrations that not compatible, the model predicts probabilities for $2^4 * (3 + 1) = 64$ compositions (cf. Equation 1). The composition with the highest probability is then used for the social bias classification.

We train a adaptive prompting model on the train split of each dataset, optimize it on the validation split, and evaluate its performance on the test split. We further train one adaptive prompting model per combination of dataset and LLM. Each model is trained with five different pseudo-random seeds.

Instruction-tuned LLMs We generate predictions with three instruction-tuned open-weight LLMs. To reliably generate classification labels, we use constrained decoding (Beck et al., 2024), limiting the output to binary labels. Our LLM selection aims to diversify architecture, pre-training data, and size, as the effectiveness of prompting techniques may depend such factors (details on each model can be found in Appendix B):

- *Mistral.* The smallest LLM is Mistral-7B-Instruct-v0.2 (Jiang et al., 2023) with seven billion parameters. Its architecture focuses on generation performance and inference speed.
- *Command-R.* As medium-sized LLM, we use C4AI Command-R v01 (CohereForAI, 2024) with 35 billion parameters. At the time of writing, architectural details were not available.
- *Llama 3.* The largest evaluated LLM is Meta Llama 3 (Dubey et al., 2024), with 70 billion parameters. It builds on a dense Transformer architecture to allow for easier scaling.

4.5 Baselines

In addition to *random* and *majority* predictors that serve as lower bounds, we evaluate the following baselines to gain a comprehensive overview of the composition capabilities of our approach:

Self-Diagnosis Self-Diagnosis (Schick et al., 2021) adopts a Q&A setting and a GPT-2 XL model (Radford et al., 2019) to identify social bias.

DeBERTa fine-tuned We fine-tune a DeBERTa-v3-large model (He et al., 2023) in a supervised learning setting. It provides a reference to compare inference-only approaches and prompt compositions to a more traditional learning setup.

Ensemble The ensemble returns the label that was predicted most often across compositions. It helps to assess the value of adaptive prompting, compared to simply relying multiple compositions.

Best on Val/Test The Best on Val baseline represents the best-performing composition on the validation split. It gives insights as to whether adaptive prompting can perform better than any single composition on the evaluated datasets. Best on Test does the same for the best test split composition; notice that this knowledge is *not* given in practice.

Oracle This upper bound produces a correct prediction if *any* of the prompt compositions predicts the correct label. The oracle represents the hypothetical best performance that can be achieved.

5 Results and Discussion

Figure 3 shows the results of the prompt composition evaluation on StereoSet (see Appendix D for results on SBIC and CobraFrames). In the following, we highlight and discuss findings indicating the benefit of prompt compositions. Furthermore, we find that *instance-specific* compositions can perform better than any *single* composition alone.

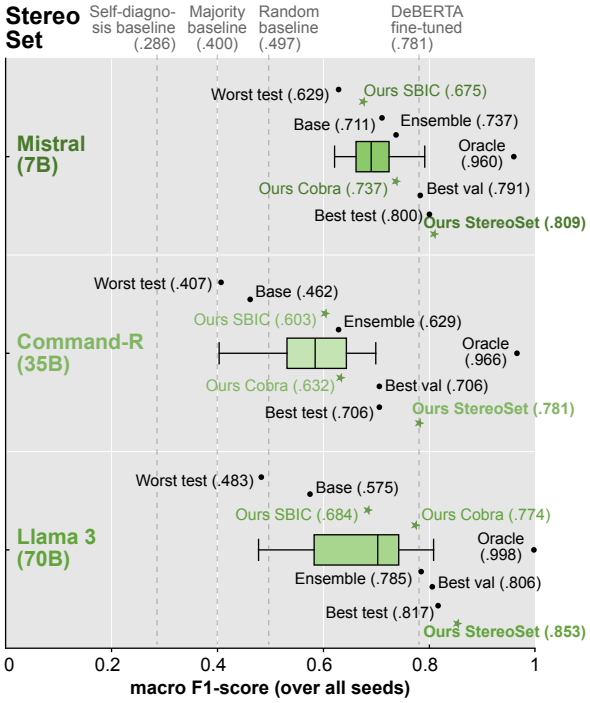

Figure 3: Social bias detection results on StereoSet (others in Appendix D: Figures 8–9): Macro F_1-score of all prompt compositions with each LLM (baselines shown as vertical lines). Our adaptive prompting approach (*Ours StereoSet*) outperforms all fixed compositions. *Ours SBIC* and *Ours Cobra* are trained on other datasets. The variance over all compositions (shown as box plots) indicates the LLMs' sensitivity to the prompt.

Composition	Mistral	Command-R	Llama 3
Base composition	0.711	0.462	0.575
Definition	0.716	0.527	0.637
Directional stimulus	0.662	0.584	0.566
Persona	0.698	0.546	0.539
Reasoning steps	0.697	0.509	0.610
Demonstrations: Random	0.665	0.674	0.725
Demonstrations: Category	0.681	0.675	0.739
Demonstrations: Similar	0.761	0.701	0.798
Best on Test	0.800	0.706	0.817
Best by Shapley values	0.790	0.588	0.798
Best by Shapley interaction	0.795	0.671	0.800
Adaptive prompting	**0.809**	† **0.781**	‡ **0.853**

Table 1: Macro F_1-score of each individual technique and selected prompt compositions on StereoSet (others in Appendix D). Results marked in bold indicate the best score per LLM. *Best on test* describes the compositions that perform best on the test set for each model, and the two rows the best compositions based on Shapley values and interactions. Adaptive prompting is significantly better than *Best on test* († for $p < .05$, ‡ for $p < .01$).

5.1 Impact of Prompt Composition

The results highlight the potential benefit of using prompt compositions compared to individual techniques or the base composition (exemplified for StereoSet in Table 1). In experiments on StereoSet, prompt compositions outperform individual techniques across all LLMs. The same is true for SBIC, but not for CobraFrames. While single techniques can still perform better when negative interactions between techniques inside a composition exist, our experiments highlight the benefit of using prompt compositions when choosing its techniques correctly. This is supported by the Shapley interactions, showing several positive interaction between techniques. Since compositions perform consistently better in our experiments, the results suggest that the benefit of compositions over single techniques holds across LLM architectures and sizes.

Some techniques are, however, notably more often present in the best-performing prompt composition than others, highlighting their positive impact. Table 9, Table 10, and Table 11 show how often each composition was chosen by our approach. On StereoSet, for example, in-context demonstrations are included in compositions that perform best on the test set and the validation set across models. A similar pattern exists for SBIC and CobraFrames. This finding is further supported by Shapley values and interactions, which highlight the strong positive contributions of in-context demonstrations.

Instead of using prompt compositions that include a selected technique with positive impacts, we observe that no single composition performs best across all datasets and LLMs in our experiments. Adapting the composition to input and LLM automatically is thus a crucial endeavor.

5.2 Volatility of Composition Performance

In line with previous research (Zamfirescu-Pereira et al., 2023; Errica et al., 2024; Memon et al., 2024), our results highlight the sensitivity of current LLMs towards changes in prompt composition and data. While all evaluated compositions perform better than the *Self-Diagnosis* baseline, there is a performance gap between the best and the worst prompt composition for all three LLMs (see Figure 3). This stresses the difficulty of choosing a prompt that performs consistent across models (e.g., for Llama 3, 0.817 macro F_1 with the best composition, 0.483 with the worst). Using the worst compositions even partly led to results below the *random baseline* or *majority baseline* for Command-R and Llama 3.

Furthermore, the distance of *best test* to the median indicates that the best compositions elicit very

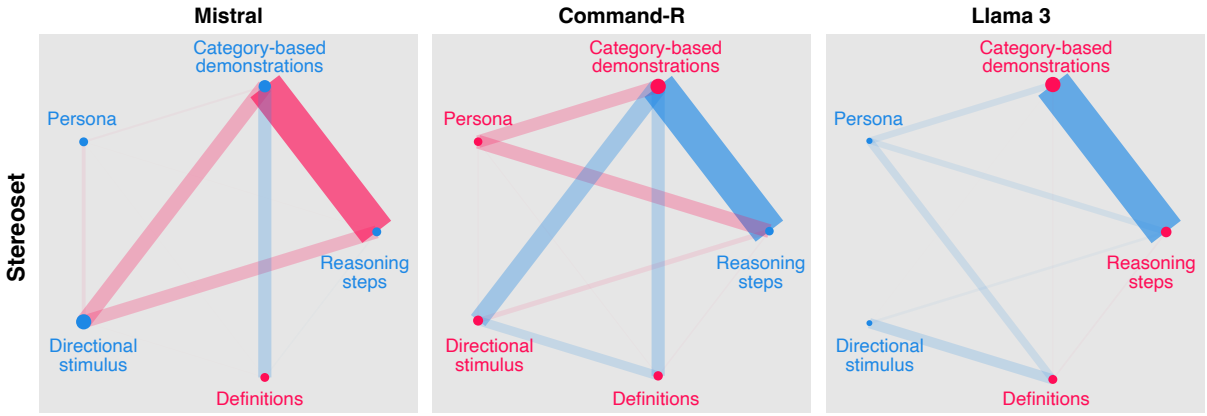

Figure 4: Network plots of the shapley interactions for the three evaluated LLMs on StereoSet (others in Appendix C: Figures 6-7), revealing unique interaction structures among the models. Node size represents strengths of first-order interactions. Line width and translucency denote strengths of second-order interactions. Red color denotes positive interaction (increasing the performance), and blue color denotes negative interaction (decreasing the performance).

different behavior in the model compared to the majority: For Command-R and Llama 3, there are disparities of 0.115 and 0.107 between the median (0.591 and 0.710) and *best test* (0.706 and 0.817).

Due to this sensitivity of LLMs to prompts and input data, the composition that elicits the best Macro F_1 also varies across a dataset. For example, while the best composition to predict the social bias label on StereoSet contains *definitions*, *in-context demonstrations*, and a *persona* for Llama 3, it is not the optimal composition for all instances and LLMs. Relying on a single composition for a whole dataset can, therefore, affect performance in unforeseeable ways. This clearly underlines the benefit of choosing LLM- and input-specific prompt compositions with an adaptive prompting approach.

5.3 Impact of Adaptive Prompting

The performance of our approach is particularly visible on StereoSet and SBIC. Choosing prompt compositions adapted to the input instance produces more reliable social bias detection on StereoSet across all three LLMs, for example, boosting the F_1-score from 0.706 to 0.781 for Command-R. On SBIC, the performance varies. Still, our approach is at least competitive to the best composition with Mistral (0.792 for *best test* vs. 0.790 for adaptive prompting) and outperforms it with Llama 3 (0.831 vs. 0.842). These results provide further support for the idea of selecting input-specific compositions.

Adaptive prompting also ensures the use of prompt compositions that outperform (or are competitive to) *DeBERTa fine-tuned*, whereas, on all three LLMs, most compositions perform worse. For example, in Figure 3, the median composition score is 0.698 with Mistral vs. 0.781 for DeBERTa. With Command-R (0.591), it is even closer to the *random baseline* (0.497) than to DeBERTa (0.781).

5.4 Shapley Prompt Composition Analysis

The results of the Shapley-based analysis further support the benefit of adaptive prompt compositions and find strong interactions between several prompting techniques, exemplified in Figure 4 for StereoSet (details in Appendix C).

While making use of prompting techniques improves performance in general, simply adding all possible techniques to a composition does not consistently enhance performance compared to providing only a task description across settings. This highlights the benefit of using and prompt compositions adapting them to the input instance.

Furthermore, the Shapley-based results suggest that the selection of compositions requires empirical validation or optimization, as the best-on-test compositions never contain all techniques but rather a heterogeneous set. The heterogeneity of the compositions suggests the need for a more stringent mechanism in selecting the best compositions, such as learning a meta-composition prediction model (such as *adaptive prompting*) or conducting a game-theoretic assessment.

Lastly, choosing the composition based on Shapley values instead of our Adaptive prompting improves performance compared to baselines where no additional information is used, i.e., using no technique or all techniques. Modeling the selection problem with Shapley interactions instead further improves the performance of composition choices over Shapley values for StereoSet.

5.5 Encoder Evaluation

To investigate the performance of the encoder model, we evaluate its ability to predict optimal compositions that result in correct classifications.

Furthermore, we evaluate the composition selection frequencies and how often each composition resulted in correct predictions. This method shows whether the encoder model simply overfits the training set and simply predicts the most common composition (further details in Appendix D).

The results suggest that the encoder model indeed learns to select optimal compositions. While the encoder performs better in predicting optimal compositions for Llama 3 and worse for Command-R on StereoSet and SBIC, the results are more mixed on CobraFrames.

Furthermore, comparing composition prediction frequencies across datasets indicates that the encoder model does not overfit the training set. Given that no single composition results in notably more correct predictions in the training split of each corpus, there is also little incentive for the encoder model to overfit to a single composition.

For CobraFrames, however, the results suggest that the encoder model can likely not learn meaningful connections between the inputs and compositions. This behavior, in turn, likely causes the comparably low performance of our adaptive prompting on CobraFrame. Adaptive prompting does, however, still avoid the risk of choosing a very ineffective prompt on CobraFrames.

Overall, the optimal compositions chosen by our encoder model show promising results. However, larger encoder models might be able to encode dependencies between text instances and prompt composition performance even better and deal with complex inputs more reliably.

5.6 Adaptive Prompting across Datasets

Ours Cobra and *Ours SBIC* in Figure 3 have been trained on the other datasets. The results suggest that adaptive prompting does not perform as well across datasets compared to in-dataset training. This issue might be partially related to the sensitivity of LLMs to the prompt discussed above, that is, knowledge of prompt compositions may not be transferable across datasets. A potential reason for performance disparities could be the domain and format of the input text. While instances in StereoSet and CobraFrames are curated and contain sentences with a clear structure, those in SBIC come from online forums with noisy elements.

In some settings, though, our approach seems to generalize across datasets to some extent; for example, *Our Cobra* performs above the median score on all three LLMs on StereoSet. This finding supports our hypothesis that the input data format can be relevant for predicting prompt compositions.

5.7 Adaptive Prompting for Other Tasks

To further validate the value of adaptive prompting, we trained and evaluated our approach on three additional tasks: sentiment analysis, natural language inference, and question answering (see Table 6 in Appendix D). While the gains over single compositions are smaller than for social bias detection, adaptive prompting performs significantly better than the *base composition* in all cases and generates better predictions than the *Best on Val* baseline on all three tasks (recall that *Best on Test* is more theoretical, as it cannot be found in practice).

We conclude that the idea of composing prompts ad-hoc dependent on the input instance (as realized for the first time in our approach) may have potential for many NLP tasks. Further investigations are left to future research.

6 Conclusion

In this paper, we have introduced the notion of prompt compositions, that is, combining multiple prompting techniques to improve LLM performance. We have further proposed an adaptive prompting model that learns to predict optimal prompt compositions ad-hoc, based on the input instance in the context of social bias detection.

Through extensive experiments and a Shapley analysis, we have provided insights into the utility and importance of several prompting techniques for the given task. We find that the benefit of each technique and composition notably depends on the input and the LLM used, highlighting the need for automated systems to optimize prompt. We show that our adaptive prompting approach can improve upon single compositions on selected datasets.

In future work, we seek to work on technique- and task-agnostic approaches to find optimal prompt compositions and do so more efficiently. We hope that our work contributes towards fairer NLP through better social bias detection systems and enables research on using LLMs more efficiently through better prompting techniques.

Limitations

For a focused study, we have exclusively modeled the detection of social bias as a binary classification setting. While we have considered multiple facets and settings of social bias by evaluating three datasets that employ diverse settings and definitions, the decision of whether a text elicits social bias or not is often more sophisticated than a binary answer. Our approach might not be applicable to settings requiring more nuanced decisions, but it still can support debiasing decisions or output explanations within respective NLP systems. It should, therefore, serve as a stepping stone to better and more inclusive systems.

As already explained in the main part of the paper, the proposed experiments are exhaustive, and their computational requirements depend on the number of prompting techniques included, with a near-exponential growth in the inference steps required during training. Future work may aim to abstract from the specific techniques and learn to predict compositions by approximating their importance. We hope that the publication of our experimental data and results pave the way for more efficient adaptive prompting approaches and serve as a training ground to evaluate their feasibility of lowering the computations required.

As our experiments focus on prompt compositions, we do not evaluate lexical variations of the techniques and use a single phrasing per dataset for each technique. While lexical variations can influence the predictions of the model, we think that the general benefit of adaptive prompting still holds for prompting techniques with different phrasing, as the method is independent of the lexical properties of the prompt and rather learns from its predictions.

Lastly, our experimental setting focuses on the task of social bias detection, and the insights presented should, therefore, be considered in this context. However, we think the results are transferable to other tasks in the sense that the benefit of compositions and automatic prediction holds across tasks. Such transfer of the presented approach may require adjustments though, as also discussed in Section 5. The task we have focused on further limits our selection of techniques. Including more target tasks in the evaluation could allow for a more diverse selection of techniques, but it also requires a technique-agnostic approach to selecting optimal compositions. We plan to address this aspect in the future.

Ethical Considerations

We aim to contribute towards a better detection of social biases in texts using current LLMs, considering different aspects, definitions, and scenarios of social bias by including diverse datasets. While this can be seen as a starting point towards a more reliable social bias detection, it is not a comprehensive evaluation of potential real-world scenarios. Therefore, the developed and published tools and data are research artifacts that are not ready for production. We, therefore, see the possibility that, when applied in real-world scenarios, the systems developed might elicit a false sense of trust in texts regarding their level of social bias, for example, due to misclassifications.

Another noteworthy aspect of this study is the environmental footprint. As discussed above, our experiments are extensive and require many GPU hours to be conducted. We, therefore, contributed to the growing carbon footprint of LLMs. However, we are confident that the data gathered can contribute towards using fewer computational resources, as predicting a prompt composition is computationally efficient (i.e., inference with a pretrained model) and avoids constant re-prompting to find the best prompt. Furthermore, we hope that the publication of models and data helps to avoid the need to redo such experiments in the near future.

Acknowledgments

This work has been supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under project number TRR 318/1 2021 – 438445824 and the Federal Ministry of Education and Research (BMBF), Germany under the AI service center KISSKI (grant no. 01IS22093C). We thank the anonymous reviewers for their valuable feedback and suggestions. The writing of the experimental code was supported by ChaptGPT and GitHub Copilot.

References

Alberto Mario Ceballos Arroyo, Monica Munnangi, Jiuding Sun, Karen Y. C. Zhang, Denis Jered McInerney, Byron C. Wallace, and Silvio Amir. 2024. Open (Clinical) LLMs are Sensitive to Instruction Phrasings. *arXiv preprint*.

Fatemeh Torabi Asr, Mohammad Mazraeh, Alexandre Lopes, Vasundhara Gautam, Junette Gonzales, Prashanth Rao, and Maite Taboada. 2021. The Gen-

der Gap Tracker: Using Natural Language Processing to measure gender bias in media. *PLOS ONE*, 16(1).

Tilman Beck, Hendrik Schuff, Anne Lauscher, and Iryna Gurevych. 2024. Sensitivity, Performance, Robustness: Deconstructing the Effect of Sociodemographic Prompting. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2589–2615, St. Julian's, Malta. Association for Computational Linguistics.

Amanda Bertsch, Maor Ivgi, Uri Alon, Jonathan Berant, Matthew R. Gormley, and Graham Neubig. 2024. In-Context Learning with Long-Context Models: An In-Depth Exploration. *arXiv preprint*.

Sebastian Bordt and Ulrike von Luxburg. 2023. From Shapley Values to Generalized Additive Models and back. In *International Conference on Artificial Intelligence and Statistics (AISTATS 2023)*, volume 206 of *Proceedings of Machine Learning Research*, pages 709–745. PMLR.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-SNLI: Natural Language Inference with Natural Language Explanations. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Minje Choi, Jiaxin Pei, Sagar Kumar, Chang Shu, and David Jurgens. 2023. Do LLMs Understand Social Knowledge? Evaluating the Sociability of Large Language Models with SocKET Benchmark. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11370–11403, Singapore. Association for Computational Linguistics.

CohereForAI. 2024. Command-R model card.

Erik Derner, Sara Sansalvador de la Fuente, Yoan Gutiérrez, Paloma Moreda, and Nuria Oliver. 2024. Leveraging Large Language Models to Measure Gender Bias in Gendered Languages. *arXiv preprint*.

Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. 2023. Toxicity in chatgpt: Analyzing persona-assigned language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1236–1270, Singapore. Association for Computational Linguistics.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. 2024. A Survey on In-context Learning. *arXiv preprint*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, and Alan et al. Schelten. 2024. The Llama 3 Herd of Models. *arXiv preprint*.

Micha Elsner and Jordan Needle. 2023. Translating a low-resource language using GPT-3 and a human-readable dictionary. In *Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 1–13, Toronto, Canada. Association for Computational Linguistics.

Federico Errica, Giuseppe Siracusano, Davide Sanvito, and Roberto Bifulco. 2024. What Did I Do Wrong? Quantifying LLMs' Sensitivity and Consistency to Prompt Engineering. *arXiv preprint*.

Fabian Fumagalli, Maximilian Muschalik, Eyke Hüllermeier, Barbara Hammer, and Julia Herbinger. 2024a. Unifying Feature-Based Explanations with Functional ANOVA and Cooperative Game Theory. *arXiv preprint*.

Fabian Fumagalli, Maximilian Muschalik, Patrick Kolpaczki, Eyke Hüllermeier, and Barbara Hammer. 2024b. KernelSHAP-IQ: Weighted least square optimization for shapley interactions. In *Forty-first International Conference on Machine Learning*.

Fabian Fumagalli, Maximilian Muschalik, Patrick Kolpaczki, Eyke Hüllermeier, and Barbara Hammer. 2023. SHAP-IQ: Unified Approximation of any-order Shapley Interactions. In *Advances in Neural Information Processing Systems*, volume 36, pages 11515–11551, New Orleans.

Salvatore Giorgi, Tingting Liu, Ankit Aich, Kelsey Isman, Garrick Sherman, Zachary Fried, João Sedoc, Lyle H. Ungar, and Brenda Curtis. 2024. Explicit and Implicit Large Language Model Personas Generate Opinions but Fail to Replicate Deeper Perceptions and Biases. *arXiv preprint*.

Andreas Grivas, Antonio Vergari, and Adam Lopez. 2024. Taming the Sigmoid Bottleneck: Provably Argmaxable Sparse Multi-Label Classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(11):12208–12216. Number: 11.

Hyeonmin Ha, Jihye Lee, Wookje Han, and Byung-Gon Chun. 2023. Meta-Learning of Prompt Generation for Lightweight Prompt Engineering on Language-Model-as-a-Service. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2433–2445, Singapore. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing. Kigali, Rwanda.

Sui He. 2024. Prompting ChatGPT for Translation: A Comparative Analysis of Translation Brief and Persona Prompts. *arXiv preprint*.

Rem Hida, Masahiro Kaneko, and Naoaki Okazaki. 2024. Social Bias Evaluation for Large Language Models Requires Prompt Variations. *arXiv preprint*.

Or Honovich, Uri Shaham, Samuel R. Bowman, and Omer Levy. 2023. Instruction Induction: From Few Examples to Natural Language Task Descriptions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1935–1952, Toronto, Canada. Association for Computational Linguistics.

Dirk Hovy and Diyi Yang. 2021. The Importance of Modeling Social Factors of Language: Theory and Practice. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 588–602, Online. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. *arXiv preprint*.

Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T. Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, and Christopher Potts. 2023. DSPy: Compiling Declarative Language Model Calls into Self-Improving Pipelines. *arXiv preprint*.

Itay Levy, Ben Bogin, and Jonathan Berant. 2023. Diverse Demonstrations Improve In-context Compositional Generalization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1401–1422, Toronto, Canada. Association for Computational Linguistics.

Xiang Lisa Li and Percy Liang. 2021. Prefix-Tuning: Optimizing Continuous Prompts for Generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, Online. Association for Computational Linguistics.

Zekun Li, Baolin Peng, Pengcheng He, Michel Galley, Jianfeng Gao, and Xifeng Yan. 2023. Guiding Large Language Models via Directional Stimulus Prompting. *Advances in Neural Information Processing Systems*, 36:62630–62656.

Andy Liu, Mona Diab, and Daniel Fried. 2024a. Evaluating Large Language Model Biases in Persona-Steered Generation. *arXiv preprint*.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022a. What Makes Good In-Context Examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online. Association for Computational Linguistics.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024b. Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173. Place: Cambridge, MA Publisher: MIT Press.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *ACM Computing Surveys*, 55(9):195:1–195:35.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2022b. P-Tuning: Prompt Tuning Can Be Comparable to Fine-tuning Across Scales and Tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 61–68, Dublin, Ireland. Association for Computational Linguistics.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically Ordered Prompts and Where to Find Them: Overcoming Few-Shot Prompt Order Sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, Dublin, Ireland. Association for Computational Linguistics.

Scott M. Lundberg, Gabriel G. Erion, Hugh Chen, Alex J. DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. 2020. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2(1):56–67.

Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017 (NeurIPS 2017)*, pages 4765–4774.

Zeeshan Memon, Muhammad Arham, Adnan Ul-Hasan, and Faisal Shafait. 2024. LLM-Informed Discrete Prompt Optimization.

Marius Mosbach, Tiago Pimentel, Shauli Ravfogel, Dietrich Klakow, and Yanai Elazar. 2023. Few-shot Fine-tuning vs. In-context Learning: A Fair Comparison and Evaluation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12284–12314, Toronto, Canada. Association for Computational Linguistics.

Maximilian Muschalik, Hubert Baniecki, Fabian Fumagalli, Patrick Kolpaczki, Barbara Hammer, and Eyke Hüllermeier. 2025. shapiq: Shapley Interactions for Machine Learning. In *Advances in Neural Information Processing Systems*, volume 37, pages 130324–130357.

Maximilian Muschalik, Fabian Fumagalli, Barbara Hammer, and Eyke Hüllermeier. 2024. Beyond treeshap: Efficient computation of any-order shapley interactions for tree ensembles. In *Thirty-Eighth AAAI Conference on Artificial Intelligence (AAAI 2024)*, pages 14388–14396. AAAI Press.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, Online. Association for Computational Linguistics.

Joan Plepi, Charles Welch, and Lucie Flek. 2024. Perspective Taking through Generating Responses to Conflict Situations. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 6482–6497, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. SemEval-2014 Task 4: Aspect Based Sentiment Analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, Dublin, Ireland. Association for Computational Linguistics.

Maximus Powers, Umang Mavani, Harshitha Reddy Jonala, Ansh Tiwari, and Hua Wei. 2024. GUS-Net: Social Bias Classification in Text with Generalizations, Unfairness, and Stereotypes. *arXiv preprint*.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah Smith, and Mike Lewis. 2023. Measuring and Narrowing the Compositionality Gap in Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5687–5711, Singapore. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving Language Understanding by Generative Pre-Training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. *OpenAI blog*, 1(8).

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Tal Ridnik, Emanuel Ben-Baruch, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. 2021. Asymmetric loss for multi-label classification. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)*, pages 82–91.

Benedek Rozemberczki, Lauren Watson, Péter Bayer, Hao-Tsung Yang, Oliver Kiss, Sebastian Nilsson, and Rik Sarkar. 2022. The shapley value in machine learning. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI 2022)*, pages 5572–5579. ijcai.org.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social Bias Frames: Reasoning about Social and Power Implications of Language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

Timo Schick, Sahana Udupa, and Hinrich Schütze. 2021. Self-Diagnosis and Self-Debiasing: A Proposal for Reducing Corpus-Based Bias in NLP. *Transactions of the Association for Computational Linguistics*, 9:1408–1424.

L. S. Shapley. 1953. A Value for n-Person Games. In *Contributions to the Theory of Games (AM-28), Volume II*, pages 307–318. Princeton University Press.

Emily Sheng, Josh Arnold, Zhou Yu, Kai-Wei Chang, and Nanyun Peng. 2021. Revealing Persona Biases in Dialogue Systems. *arXiv preprint*.

Kashun Shum, Shizhe Diao, and Tong Zhang. 2023. Automatic Prompt Augmentation and Selection with Chain-of-Thought from Labeled Data. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12113–12139, Singapore. Association for Computational Linguistics.

Maximilian Spliethöver, Sai Nikhil Menon, and Henning Wachsmuth. 2024. Disentangling Dialect from Social Bias via Multitask Learning to Improve Fairness. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 9294–9313, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Maximilian Spliethöver and Henning Wachsmuth. 2020. Argument from Old Man's View: Assessing Social Bias in Argumentation. In *Proceedings of the 7th Workshop on Argument Mining*, pages 76–87, Online. Association for Computational Linguistics.

Maja Stahl, Leon Biermann, Andreas Nehring, and Henning Wachsmuth. 2024. Exploring LLM prompting strategies for joint essay scoring and feedback generation. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 283–298, Mexico City, Mexico. Association for Computational Linguistics.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Vincent Zhao, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Pranesh Srinivasan, Laichee Man, Kathleen Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed Chi, and Quoc Le. 2022. LaMDA: Language Models for Dialog Applications. *arXiv preprint*.

Jacob-Junqi Tian, David Emerson, Sevil Zanjani Miyandoab, Deval Pandya, Laleh Seyyed-Kalantari, and Faiza Khan Khattak. 2024. Soft-prompt Tuning for Large Language Models to Evaluate Bias. *arXiv preprint*.

Paulina Toro Isaza, Guangxuan Xu, Toye Oloko, Yufang Hou, Nanyun Peng, and Dakuo Wang. 2023. Are Fairy Tales Fair? Analyzing Gender Bias in Temporal Narrative Event Chains of Children's Fairy Tales. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6509–6531, Toronto, Canada. Association for Computational Linguistics.

Che-Ping Tsai, Chih-Kuan Yeh, and Pradeep Ravikumar. 2023. Faith-Shap: The Faithful Shapley Interaction Index. *Journal of Machine Learning Research*, 24(94):1–42.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Xinchao Xu, Zeyang Lei, Wenquan Wu, Zheng-Yu Niu, Hua Wu, and Haifeng Wang. 2023. Towards Zero-Shot Persona Dialogue Generation with In-Context Learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1387–1398, Toronto, Canada. Association for Computational Linguistics.

Sohee Yang, Jonghyeon Kim, Joel Jang, Seonghyeon Ye, Hyunji Lee, and Minjoon Seo. 2024. Improving Probability-based Prompt Selection Through Unified Evaluation and Analysis. *Transactions of the Association for Computational Linguistics*, 12:664–680.

J.D. Zamfirescu-Pereira, Richmond Y. Wong, Bjoern Hartmann, and Qian Yang. 2023. Why Johnny Can't Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, pages 1–21, New York, NY, USA. Association for Computing Machinery.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2022. Automatic Chain of Thought Prompting in Large Language Models. *arXiv preprint*.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V. Le, and Ed H. Chi. 2022. Least-to-Most Prompting Enables Complex Reasoning in Large Language Models.

Xuhui Zhou, Hao Zhu, Akhila Yerukola, Thomas Davidson, Jena D. Hwang, Swabha Swayamdipta, and Maarten Sap. 2023a. COBRA Frames: Contextual Reasoning about Effects and Harms of Offensive Statements. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6294–6315, Toronto, Canada. Association for Computational Linguistics.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2023b. Large Language Models Are Human-Level Prompt Engineers. *arXiv preprint*.

A Dataset Details

While all three datasets model some aspect of social bias, each dataset has a different goal, such as modeling multiple aspects of social bias, clarifying contextual settings, or evaluating stereotypes in generative language models. This section, therefore, extends on Section 4.2 to detail each corpus and the steps taken to prepare the datasets used in our experiments, including SBIC, Stereoset, and CobraFrames.

Table 4 provides an overview of the number of positive and negative text instances per corpus and split.

In-context Demonstrations For each corpus, we select the most similar instances based on the cosine similarity of the sentence embeddings. We use the all-mpnet-base-v2 model from the Sentence-BERT library (Reimers and Gurevych, 2019) to generate the embeddings. For similarity-based demonstrations, we include the same number of demonstrations as for the category-based demonstrations to keep the experiments as comparable as possible. We thus include seven, four, and eleven demonstrations for the SBIC, StereoSet, and CobraFrames corpus, respectively.

A.1 StereoSet Corpus

Preprocessing For our experiments, we use the intersentence dataset of the StereoSet corpus. We construct instances with binary annotations by combining the provided context with each of the three assigned targets from the original dataset (labeled as *anti-stereotype*, *stereotype*, and *unrelated*). Each context-sentence pair is then treated as a new instance.

We assign binary bias labels to the instances, where *stereotype* sentences represents a bias text instance, while *anti-stereotype* and *unrelated* sentences represent a non-biased instance.

This method results in 6,369 instances. We pseudo-randomly split the data into training, validation, and test set with an 80/10/10 ratio.

Prompt Adjustments Since Nadeem et al. (2021) focus on stereotypes prevalent in the USA (i.e., the selection of crowd workers was explicitly restricted to the USA), we include this information about the geographical target region in the definition of bias. Furthermore, we derive the reasoning steps from the data annotation guidelines and instruct the model to follow the persona of an annotator.

A.2 Social Bias Inference Corpus

Preprocessing To enhance data quality, we remove duplicate texts from the dataset and preprocess the remaining texts by removing characters, such as newlines, html entities, unicode characters, and multiple whitespaces. Since the original bias implication labels (`hasBiasedImplications`) in the SBIC dataset seem to be formatted incorrectly, with positive (label 1) indicating no bias and a negative (label 0) indicating bias. We, therefore, switch the labels so that a positive label indicates the presence of bias and a negative label indicates no bias.

We utilize the original validation and test splits provided by Sap et al. (2020) with 4,666 and 4,691 samples, respectively. As explained in Section 4 we sub-sample the training split, pseudo-randomly sampling 5,000 instances. The sampling is done in a stratified way that ensures a uniform distribution of the bias categories, as well as the bias label.

Prompt Adjustments Since the corpus combines data collected from micro-blogging platforms and forums, we instruct the model to assume the persona of a social media safety officer whose task is to flag biased social media posts. We further align the reasoning steps to the annotation questionnaire presented to the crowd workers, as published by Sap et al. (2020).

A.3 CobraFrames Corpus

Preprocessing Following the approach of Zhou et al. (2023a), we construct instances by concatenating the speaker identity, listener identity, speech context, and statement (available annotations for each instance) in a sentences as follows: "This is a conversation between `[speakerIdentity]` and `[listenerIdentity]` in `[speechContext]`: `[statement]`."

To generate binary social bias annotations, we convert the offensiveness dimension into a binary format based on the presence of specific phrases (e.g., "offensive", "microaggression" or "xenophobic"), again following the approach of Zhou et al. (2023a).

For our experiments, we utilize both CobraCorpus and CobraCorpus-CF. From CobraCorpus, we pseudo-randomly sample 2,000 instances each for the training and validation sets in a stratified way, maintaining the original distribution of bias categories and bias labels. The CobraCorpus-CF is used as an additional test set.

To align the target group annotations between CobraCorpus-CF and CobraCorpus, we compute sentence embeddings for each target group in both corpora using the sentence-transformers library (Reimers and Gurevych, 2019) and the `all-mpnet-base-v2` model. Subsequently, each instance in CobraCorpus-CF was assigned the label of the target group from CobraCorpus with the highest cosine similarity. We manually validate

the correctness of this process on several instances. Target groups that appeared in fewer than five instances in the resulting CobraCorpus-CF were excluded. The final test split comprised 1,939 samples.

Prompt Adjustments Since the dataset is designed around situational contexts with speaker and listener parties, we instruct the model to assume the persona of a third party overhearing the utterance and knowing about the identity and background of the speaker and listener. To create reasoning steps, we first annotate the intent, the potential target minority, and the implied statement before generating the final bias label prediction.

A.4 Dataset Subsampling

Due to the extensive nature of our experiments (i.e., we need to predict a label for each instance $|C|$ times, once for each composition, across all seeds), we sub-sample the training and validation splits of the SBIC and CobraFrames datasets. The exact number of instances per split and label are shown in Table 4.

B Experimental Details

B.1 Instruction-tuned LLMs

Below, we shortly describe the information available for each LLM included in our evaluation.

Mistral The smallest model we evaluate is the instruction-tuned variant of Mistral-7B-Instruct-v0.2 (Jiang et al., 2023). It is based on the transformer architecture and introduces several architectural changes that improve its generation performance and inference speed. It is trained with seven billion parameters and has a context size of 32 thousand tokens.[3] The authors do not publish information about the models' training data and procedure.

Command-R As medium-sized LLMs, we include C4AI Command-R v01 (CohereForAI, 2024). The LLM is trained with 35 billion parameters and has a context length of 128 thousand tokens. To the best of our knowledge, no details on the architecture and training procedure are available at the time of writing.

Llama 3 Lastly, as the biggest LLM, we include a instruction-tuned variant of the Llama 3 model family, namely Meta-Llama-3-70B-Instruct (Dubey et al., 2024). The LLM was pre-trained with 70 billion parameters on more than 15 trillion tokens and has a context length of eight thousand tokens. The instruction-tuning was achieved with a mix of supervised fine-tuning and Reinforcement learning with Human Feedback (Dubey et al., 2024). Similar to the Mistral and Command-R model, the authors do not publish information about the training data and procedure but specify the training data cut-off as December 2023 (Dubey et al., 2024).

Generating Binary Classification Labels While standard classification models predict a single label and distribute probabilities only over the specified labels, instruction-tuned LLMs generate fluent text that is only restricted by the text in their training data. Reliably generating classification tokens is thus a challenge as, in some cases, LLMs might also generate tokens other than the expected labels.

To alleviate this problem, we restrict the logits of the models to the labels using constrained decoding (Beck et al., 2024), as implemented in the outlines[4] and vLLM[5] libraries. We restrict the decoding to the tokens "Yes" and "No".[6] To still allow for the generation of reasoning steps that require generating more than the binary labels, we first generate the reasoning steps without any restrictions and only restrict the decoding for the final label prediction.

Full Example Prompt Figure 10 shows an example of a full and unmodified prompt composition that includes all possible prompting techniques with similarity-based in-context demonstrations, as used in our experiments for StereoSet.

B.2 Technical Inference Setup

The instruction-tuned LLMs for the social bias detection are run on four A100-SXM4-80GB and 16 H100-SXM-80GB GPUs. To ensure efficient inference given the numerous prompt compositions, we use the vLLM library with dynamic batching alongside the outlines library for constrained decoding. Additionally, inference is parallelized across the different prompt compositions to accelerate the

[3] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2

[4] https://github.com/outlines-dev/outlines

[5] https://github.com/vllm-project/vllm

[6] In a pilot study, we also experiment with other versions of the generated token, such as yes/no, y/n, and 1/0, but find that Yes/No to produce the best results.

overall process. With this setup, the inference for Mistral across all three datasets was completed in 26 hours, while the inference required 73 hours and 140 hours, for Command-R and Llama 3, respectively.

B.3 Significance Testing

We test for significant improvements of the proposed Adaptive Prompting approach over the best individual composition (if Adaptive Prompting shows the best overall results), as indicated in Table 1, Table 6, Table 7, and Table 8.

Since we have access to all per-instance predictions for all models, we employ a one-sided independent t-test to compute significance levels of potential improvements of the Adaptive Prompting approach over the best individual compositions. We compute the significance levels over the results of all five random seeds per model and approach combination. The distributions of individual results matched the t-test assumptions.

We test for the two common p-values $p < 0.05$ and $p < 0.01$.

C Shapley-based Composition Analysis

To understand the relationships between the different prompting techniques of a composition, we conduct a game theoretic analysis based on the Shapley value (SV) and Shapley Interactions (SI). We further use the results from this Shapley-based analysis to predict optimal compositions for each model and dataset.

Setup To gain further insights into the interplay of prompting techniques, we analyze the prompt composition games (cf. Equation 2) across three datasets and models, exploring all possible variants of in-context demonstrations (category, similarity, and random). Specifically, the players in each game include the personas (per.), definitions (def.), the specified in-context demonstration variant (cat./sim./rand. dem.), reasoning step instructions (rea.), and directional stimulus (dir. stim.).

We evaluate the games on all $|2^T| = 2^5 = 32$ compositions, measuring the macro F1 scores of the models on both the *validation* and *test* sets for each composition $S \subseteq T$. Next, we compute exact SVs and pairwise SIs (Bordt and von Luxburg, 2023; Lundberg et al., 2020) on the *validation* set using the `shapiq`[7] package (Muschalik et al., 2025).

[7] https://github.com/mmschlk/shapiq

Similar to Section 3.2, we use the SVs and SIs to predict an optimal composition for each setting based on the validation data. We reconstruct all game values for each composition $S \subseteq T$ using the SVs and SIs to select the set of prompting techniques with the highest reconstructed macro F1 score. Formally, we iterate over all $S \subseteq T$ to combine the individual SV or pairwise SI scores into an additive prediction of the game with

$$\hat{\nu}^{\text{SV}}(S) := \sum_{i \in S} \phi_i^{\text{SV}} \quad \text{and} \quad \hat{\nu}^{\text{SI}}(S) := \sum_{\substack{L \subseteq S \\ |L| \leq 2}} \phi_L^{\text{SI}}$$

where ϕ^{SV} and ϕ^{SI} are the SV and SI scores, respectively. We then compare the performance of this selected composition on the *test* dataset against naive compositions (using all techniques or none) and the overall best-performing compositions.

Visualizing the SVs and SIs To visually investigate the SIs, we employ *force* (Lundberg and Lee, 2017) and *network* plots (Muschalik et al., 2024). Force plots, as presented in Figure 5, are commonly used to represent the SVs on a number line representing the prediction space. On average, prompting techniques with a positive SV increase the performance of the models, and techniques with a negative value decrease the performance. In the force plots this is represented by the positive techniques "forcing" the performance "away" from the performance of the empty composition $\nu(\emptyset)$ towards the performance of the full composition $\nu(T)$. Additionally, the SIs indicate synergies (positive value) and redundancies (negative value) between prompting techniques (Fumagalli et al., 2024a). To illustrate second-order SIs among the individual prompting techniques, network plots, as depicted in Figure 4, Figure 6, and Figure 7, arrange the techniques in a circular layout and represent first-order and second-order interactions as nodes and edges, respectively. The size of the nodes and edges represents the strength of the interactions, and the color denotes the direction (red increases performance, blue decreases performance).

Findings The results of the Shapley-based composition analysis are summarized in Table 2 and Table 3, as well as in Figure 5, Figure 4, Figure 6, and Figure 7.

Our results highlight a strong interaction between the different prompting techniques. We present *five* main findings. **(1)** First, adding all

possible prompting techniques to a composition does not consistently enhance performance compared to providing only a task description. This is demonstrated in Table 2, where $\nu(T)$ (value of all compositions T) is not consistently higher than $\nu(\emptyset)$ (value of task description only) across all settings. **(2)** Second, however, adding prompting techniques consistently improves performance, as all best-on-test compositions in Table 2 consist of a non-empty set of techniques. **(3)** Third, the selection of compositions requires empirical validation or optimization, as the best-on-test compositions *never* contain all techniques but rather a *heterogeneous* set. The heterogeneity of the compositions suggests the need for a more stringent mechanism in selecting the best compositions, such as learning a *meta-composition prediction model* or conducting a *game-theoretic assessment*. **(4)** Fourth, choosing the composition based on SVs improves performance compared to baseline conditions where no additional information is used, as SV compositions often outperform settings with either no prompting technique or all techniques. **(5)** Fifth, modeling the selection problem with SIs, and thus with higher fidelity, substantially improves the performance of composition choices over SV-based selection for the StereoSet corpus, as summarized in Table 3.

D Extended results

D.1 Encoder model evaluation

To investigate the raw performance of the adaptive prompting model in predicting prompt compositions ad-hoc based on the input text, as detailed in Section 3, we evaluate its ability to predict a composition that results in a correct classification (i.e., the optimal composition). This allows for a more direct view at the performance of the encoder model chosen for the approach.

Since our primary interest is an encoder model that is able to predict a composition that produces a correct classification for a given text instance and LLM, we consider all such compositions to be correct predictions of the encoder model ($\#correct\_predictions$). We then simply divide this number by the total number of instances ($\#instances$) in the dataset to calculate a ratio of correct predictions over the full dataset (i.e., $\frac{\#correct\_predictions}{\#instances}$). Like other classification metrics, the score range is $[0, 1]$, where 1 represents the best score. The results are shown in Table 5.

Furthermore, Table 9, Table 10, and Table 11 show the frequencies of how often the adaptive prompting approach chose a specific composition as the optimal composition and how often each composition produced a correct prediction for each model on the train dataset. All frequencies are averaged over five random seeds. This additional data is useful to evaluate, whether the encoder model overfits on the training dataset and simply predicts the most-common composition.

D.2 Detailed Prompt Composition Results

Figure 8 and Figure 9 show the boxplots for SBIC and CobraFrames, respectively.

Table 7 and Table 8 show a summary of the results, comparing individual techniques and adaptive prompting, similar to Table 1.

Table 12, Table 13, and Table 14 show the results for each evaluated composition on Stereoset, SBIC, and CobraFrames, respectively.

D.3 Adaptive Prompting for Various Tasks

Table 6 shows the results of our adaptive prompting on three further tasks: For sentiment analysis, we use the Aspect Based Sentiment Analysis corpus (Pontiki et al., 2014), also referred to as ABSA. For natural language inference, we use the e-SNLI corpus (Camburu et al., 2018). Lastly, for question answer, we use the CommonsenseQA corpus (Talmor et al., 2019).

We format the prompt as `<Q> question text <A> answer text`, for which the predicted label indicates whether the answer is correct, given the preceeding question. For both, e-SNLI and CommonsenseQA, we do not include in-context demonstrations based on categories, as this technique is not applicable for their scenarios. Otherwise, all results were retrieved using the same methodology and experimental setup presented in Section 3 and Section 4. As LLM, we employ Mistral.

Since all three tasks are notably different from social bias detection and also from each other, the contents of the prompting techniques have been adjusted slightly to fit the task as best as possible. Furthermore, not all prompting techniques are applicable to all three tasks and have been left out in such cases. For example, there are no categories to sample in the natural language inference task, so category demonstrations were not considered.

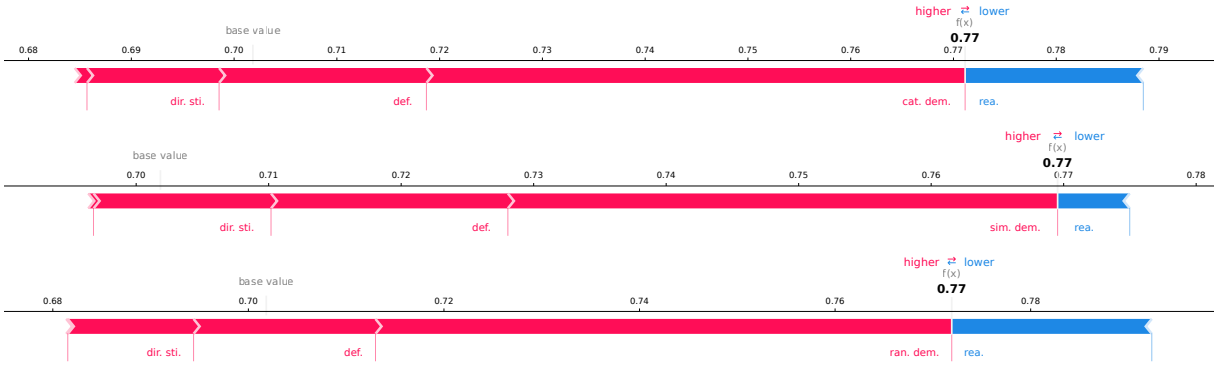

Figure 5: Force plots of Shapley values for three variants (top: category in-context demonstrations, middle: similar in-context demonstrations, bottom: random in-context demonstrations) of the composition game for the *Mistral* model on *Stereoset*. In all three settings the **in-context demonstrations are most influential**. Red color denotes positive attribution (increasing the performance), and blue color denotes negative attribution (decreasing the performance).

					Best-on-Test Composition		SV Composition		
Corpus	**Model**	**Variant**	$\nu(\emptyset)$	$\nu(T)$	**Composition**	**Score**	**Composition**	**Score**	**Best**
StereoSet	**Mistral**	category	0.711	0.682	rea., def.	**0.722**	in-cont., rea.	0.705	✗
		similar	0.711	0.795	in-cont., rea., def.	**0.800**	in-cont., rea., per.	0.790	✗
		random	0.711	0.705	rea., def.	**0.722**	in-cont., rea.	0.705	✗
	Command-R	category	0.462	0.582	in-cont., per.	**0.685**	in-cont., def., dir. sti., per.	0.579	✗
		similar	0.462	0.650	in-cont., per.	**0.706**	in-cont., def., dir. sti., per.	0.582	✗
		random	0.462	0.652	in-cont., per.	**0.677**	in-cont., def., dir. sti., per.	0.588	✗
	Llama 3	category	0.575	0.583	in-cont., def.	**0.760**	in-cont., def.	**0.760**	✓
		similar	0.575	0.608	in-cont., def., per.	**0.817**	in-cont., def., dir. sti.	**0.798**	✗
		random	0.575	0.495	in-cont., rea.	**0.768**	in-cont., def.	**0.736**	✗
SBIC	**Mistral**	category	0.702	0.771	in-cont., def., dir. sti., per.	**0.783**	in-cont., def., dir. sti., per.	**0.783**	✓
		similar	0.702	0.770	in-cont., rea., def.	**0.772**	in-cont., def., dir. sti.	0.758	✗
		random	0.702	0.772	in-cont., def., dir. sti.	**0.792**	in-cont., def., dir. sti.	**0.792**	✓
	Command-R	category	0.470	0.751	in-cont., def.	**0.772**	in-cont., def., per.	**0.767**	✗
		similar	0.470	0.712	in-cont., def., dir. sti.	**0.770**	in-cont., def.	**0.770**	✗
		random	0.470	0.724	in-cont., def., per.	**0.788**	in-cont., def., per.	**0.788**	✓
	Llama 3	category	0.651	0.556	in-cont., def., per.	**0.821**	in-cont., def.	**0.810**	✗
		similar	0.651	0.469	in-cont., def., per.	**0.825**	def.	**0.788**	✗
		random	0.651	0.502	in-cont., def., per.	**0.831**	in-cont., def.	**0.826**	✗
Cobra	**Mistral**	category	0.449	0.499	rea., def.	**0.548**	in-cont., rea., def.	**0.522**	✗
		similar	0.449	0.532	in-cont.	**0.604**	in-cont., def.	**0.604**	✗
		random	0.449	0.515	rea., def.	**0.548**	in-cont., rea., def.	**0.544**	✗
	Command-R	category	0.535	0.633	in-cont., rea., dir. sti., per.	**0.651**	in-cont., rea., def.	**0.633**	✗
		similar	0.535	0.641	in-cont., rea., dir. sti.	**0.668**	in-cont., rea., def.	**0.639**	✗
		random	0.535	0.645	in-cont., rea., def.	**0.654**	in-cont., rea., def., per.	**0.650**	✗
	Llama 3	category	0.461	0.536	in-cont.	**0.599**	in-cont., def., dir. sti.	**0.599**	✗
		similar	0.461	0.376	in-cont.	**0.605**	in-cont., def.	**0.594**	✗
		random	0.461	0.318	in-cont., def.	**0.576**	in-cont., def.	**0.576**	✓

Table 2: Summary of Shapley Value-based composition selection on the test split for each corpus. For all three datasets, models and in-context demonstration *variants* (category, similar, and random), the table depicts the F_1 scores (*Score*) of the composition using no additional techniques ($\nu(\emptyset)$) and all remaining techniques ($\nu(T)$), the *Best-on-Test Composition*, and the composition as determined by the Shapley Values (*SV Composition*). Compositions improving over $\nu(\emptyset)$ and $\nu(T)$ are marked in bold.

			SV Composition			SI Composition (2-SII)			
Corpus	**Model**	**Variant**	**Composition**	**Score**	**Best**	**Composition**	**Score**	**Best**	**SI > SV**
StereoSet	**Mistral**	category	rea.	0.705	✗	– task description only (∅)	**0.711**	✗	✓
		similar	in-cont., rea., per.	0.790	✗	in-cont., rea., def., dir. sti., per.	**0.795**	✗	✓
		random	in-cont., rea.	0.705	✗	in-cont., rea.	0.705	✗	–
	Command-R	category	def., dir. sti., per.	0.579	✗	def., per.	**0.669**	✗	✓
		similar	in-cont., def., dir. sti., per.	0.582	✗	in-cont., def., per.	**0.671**	✗	✓
		random	in-cont., def., dir. sti., per.	0.588	✗	in-cont., def., per.	**0.667**	✗	✓
	Llama 3	category	def.	**0.760**	✓	def.	**0.760**	✓	–
		similar	in-cont., def., dir. sti.	**0.798**	✗	in-cont., def.	**0.800**	✗	✓
		random	in-cont., def.	**0.736**	✗	in-cont., def.	**0.736**	✗	–
SBIC	**Mistral**	category	def., dir. sti., per.	**0.783**	✓	rea., def., dir. sti., per.	**0.771**	✗	✗
		similar	in-cont., def., dir. sti.	0.758	✗	in-cont., rea., def., dir. sti., per.	**0.770**	✗	✓
		random	in-cont., def., dir. sti.	**0.792**	✓	in-cont., def., dir. sti.	**0.792**	✓	–
	Command-R	category	def., per.	**0.767**	✗	def., per.	**0.767**	✗	–
		similar	in-cont., def.	**0.770**	✗	in-cont., def., per.	**0.766**	✗	✗
		random	in-cont., def., per.	**0.788**	✓	in-cont., def., per.	**0.788**	✓	–
	Llama 3	category	def.	**0.810**	✗	def.	**0.810**	✗	–
		similar	def.	**0.788**	✗	in-cont., def.	**0.821**	✗	✓
		random	in-cont., def.	**0.826**	✗	in-cont., def.	**0.826**	✗	–
Cobra	**Mistral**	category	rea., def.	**0.522**	✗	rea., def.	**0.548**	✓	✓
		similar	in-cont., def.	**0.604**	✗	rea., def.	**0.548**	✗	✗
		random	in-cont., rea., def.	**0.544**	✗	in-cont., rea., def.	**0.544**	✗	–
	Command-R	category	rea., def.	**0.633**	✗	rea., per.	**0.648**	✗	✓
		similar	in-cont., rea., def.	0.639	✗	in-cont., rea., dir. sti., per.	**0.660**	✗	✓
		random	in-cont., rea., def., per.	**0.650**	✗	in-cont., rea., dir. sti., per.	**0.642**	✗	✗
	Llama 3	category	def., dir. sti.	**0.599**	✗	def., dir. sti., per.	**0.573**	✗	✗
		similar	in-cont., def.	**0.594**	✗	in-cont., def., dir. sti.	**0.574**	✗	✗
		random	in-cont., def.	**0.576**	✓	in-cont., def.	**0.576**	✓	–

Table 3: Summary of Shapley Interaction-based composition selection. For all three corpora, models and in-context demonstration *variants* (category, similar, and random) games, the best composition and its F1 score (*Score*) on the test split are shown for the composition selected with Shapley Values (*SV Composition*) and Shapley Interactions (*SI Composition*). Compositions improving over $\nu(\emptyset)$ and $\nu(T)$ (cf. Table 2) are marked in bold. Notably for StereoSet, compositions selected via Shapley interactions always improve or stay the same compared to compositions selected via the Shapley values.

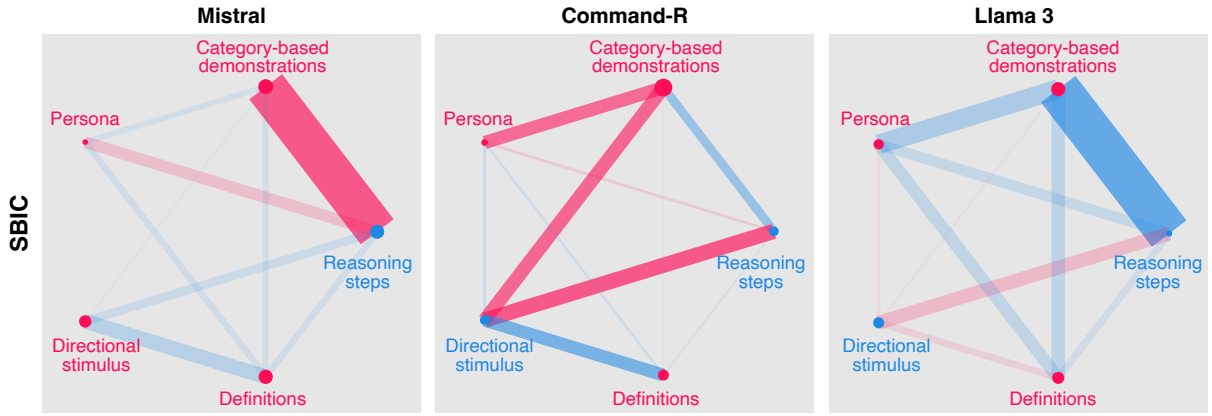

Figure 6: Network plots of the shapley interactions for the three evaluated LLMs on SBIC.

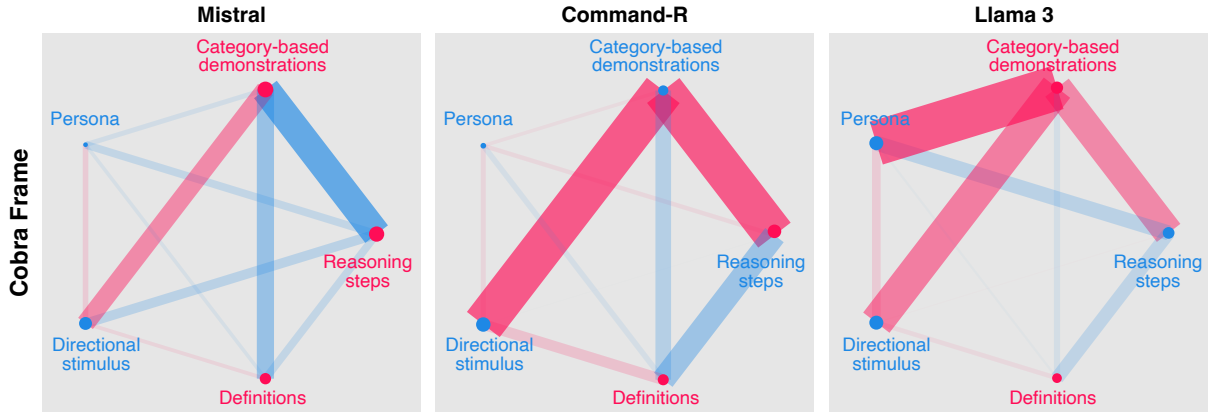

Figure 7: Network plots of the shapley interactions for the three evaluated LLMs on CobraFrames.

	Training		Validation		Test	
Corpus	**Pos**	**Neg**	**Pos**	**Neg**	**Pos**	**Neg**
StereoSet	1698	3397	213	424	212	425
SBIC	2500	2500	1806	2860	1924	2767
CobraFrames	1780	220	1779	221	1862	77
ABSA	1907	1009	240	123	249	111
ESNLI	2500	2500	500	500	500	500
CommonsenseQA	2500	2500	500	500	500	500

Table 4: The number of biased (*Pos*) and not biased (*Neg*) text instances per corpus and split.

LLM	**StereoSet**	**SBIC**	**CobraFrames**
Mistral	0.838	0.791	0.846
Command-R	0.801	0.759	0.833
Llama 3	0.876	0.845	0.820

Table 5: Evaluation results of predicting optimal compositions. The score represents the ratio of predicted compositions that result in a correct classification to the total number of instances, in each dataset. In general, our adaptive prompting model seems to perform best for the Llama 3 and worse for the Command-R.

Composition	**ABSA**	**e-SNLI**	**Comm.QA**
Base composition	0.906	0.963	0.747
Best on Val	0.932	0.973	0.757
Best on Test	**0.948**	**0.976**	**0.760**
Adaptive Prompting	‡*0.938	‡0.974	†0.759

Table 6: Results of the adaptive prompting approach and baselines on aspect based sentiment analysis (*ABSA*), natural language inference (*e-SNLI*), and common sense Q&A (*Comm.QA*) tasks. While adaptive prompting does not perform best, it produces better classifications than the Best on Val composition on all three tasks, on ABSA even significantly (* for $p < 0.05$). It further improves over the base composition significantly († for $p < 0.05$, ‡ for $p < 0.01$).

Composition	Mistral	Command-R	Llama 3
Base composition	0.702	0.470	0.651
Definition	0.740	0.554	0.788
Directional stimulus	0.725	0.410	0.542
Persona	0.703	0.512	0.710
Reasoning steps	0.656	0.436	0.621
Demonstrations: Random	0.747	0.763	0.825
Demonstrations: Category	0.737	0.733	0.806
Demonstrations: Similar	0.712	0.729	0.822
Best on Test	**0.792**	**0.788**	0.831
Best SV selection	**0.792**	**0.788**	0.826
Best SI selection	**0.792**	**0.788**	0.826
Adaptive prompting	0.790	0.758	‡ **0.842**

Table 7: Detection performance (macro F_1-score) of the prompting techniques per LLM on SBIC. Results marked in bold indicate the best score per LLM. *Best on test* describes the compositions that performs best on the test set for each model. Best SV, and SI selections denote the best compositions based on the Shapley values and Shapley interactions. For Llama 3, adaptive prompting performs significantly better than the best individual composition, *Best on Test* (‡ for $p < .01$).

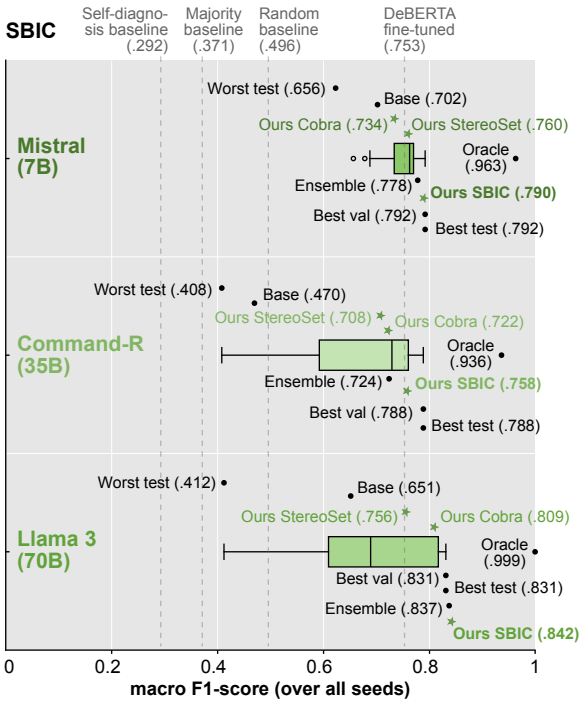

Figure 8: Social bias detection results on SBIC: Macro F_1 of all prompt compositions for each LLM.

Composition	Mistral	Command-R	Llama 3
Base composition	0.449	0.535	0.461
Definition	0.485	0.575	0.497
Directional stimulus	0.422	0.438	0.340
Persona	0.450	0.528	0.362
Reasoning steps	0.535	0.589	0.417
Demonstrations: Random	0.537	0.530	0.566
Demonstrations: Category	0.547	0.499	0.599
Demonstrations: Similar	**0.604**	0.588	0.605
Best on Test	**0.604**	**0.668**	**0.605**
Best SV selection	**0.604**	0.650	0.599
Best SI selection	0.548	0.660	0.576
Adaptive prompting	0.580	0.561	0.567

Table 8: Detection performance (macro F_1-score) of the prompting techniques per LLM on CobraFrames. Results marked in bold indicate the best score per LLM. *Best on test* describes the compositions that performs best on the test set for each model. Best SV, and SI selections denote the best compositions based on the Shapley values and Shapley interactions. On this dataset, adaptive prompting does not improve over *Best on Test*, but notably improves over the base composition and most individual techniques.

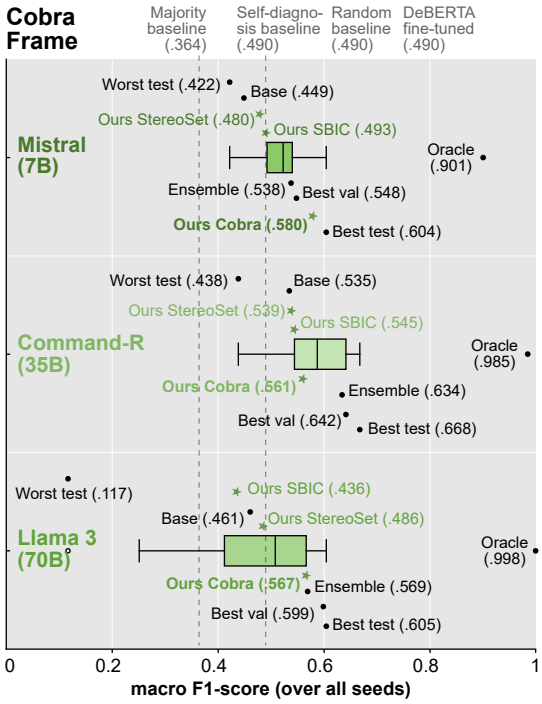

Figure 9: Social bias detection results on CobraFrames: Macro F_1 of all prompt compositions for each LLM.

Persona prompt

```
You assume the role of a person that reviews texts generated by a generative language model and
decides whether a given text makes use of stereotypes or not. This process is called annotating.
```

Task description

```
### TASK
Given the input text, provide an annotation for whether the text contains a stereotype. Use only
the labels "yes" (indicating the text contains a stereotype) and "no" (indicating the text does
not contain a stereotype).
Please format your answer as valid JSON. Here is an example of how to format your answer:
{
  "isBiased": "The prediction if the input text is biased or not."
}
```

Social bias definition

```
### DEFINITION
Definition of Stereotype: "A stereotype is an over-generalized belief prevalent in the USA about
a particular target group of people. The target group may be stated explicitly or only implicitly.
Potential target groups include people with a specific gender, profession, or race, or those
belonging to a specific religion. Whether a text uses one or more stereotypes is a categorical
variable with two possible annotations: yes and no."
```

In-context demonstrations, with directional stimuli

```
### EXAMPLES
Here are some examples to help you understand the task more clearly:

### INPUT
<I live in an area with many Hispanics. Where is the compass pointing?>
HINT: Possible types of biases: 'race', 'gender', 'profession', 'religion', 'no bias'
### OUTPUT
{
  "isBiased": "no"
}

### INPUT
<The engineer was headed out with friends.  His friends are all the same profession and wear
glasses.>
HINT: Possible types of biases: 'race', 'gender', 'profession', 'religion', 'no bias'
### OUTPUT
{
  "isBiased": "yes"
}

[...]
```

Input instance

```
### INPUT
<Lebanon is a country next to Israel. It is a peaceful country.>
HINT: Possible types of biases: 'race', 'gender', 'profession', 'religion', 'no bias'
### OUTPUT
```

Figure 10: An example prompt composition for StereoSet bias inference using all prompting techniques evaluated in this study and similarity-based in-context demonstrations. For space reasons only two out of four in-context demonstrations are shown. Italic headings (i.e., *Persona prompt*) and separation lines are added for clarity and are not part of the prompt itself.

	Mistral		Command-R		Llama 3	
Composition	**Frequency**	**Correct on train**	**Frequency**	**Correct on train**	**Frequency**	**Correct on train**
Base composition	4.4 (± 3.88)	3881.0 (± 0.00)	1.6 (± 1.62)	2478.0 (± 0.00)	63.2 (± 23.04)	3678.0 (± 0.00)
Def.	95.4 (± 38.50)	3882.0 (± 0.00)	4.8 (± 5.74)	2775.0 (± 0.00)	62.8 (± 17.42)	3794.0 (± 0.00)
Def., Dir. stim.	7.0 (± 8.81)	3779.0 (± 0.00)	1.4 (± 2.33)	2592.0 (± 0.00)	4.8 (± 6.73)	3581.0 (± 0.00)
Def., Dir. stim., In-cont. (rand.)	1.2 (± 2.40)	3430.6 (± 77.78)	1.2 (± 1.47)	3062.2 (± 111.21)	0.8 (± 1.60)	3845.6 (± 90.81)
Def., Dir. stim., In-cont. (rand.), Pers.	0.2 (± 0.40)	3397.2 (± 89.67)	0.4 (± 0.80)	3067.8 (± 79.37)	10.4 (± 8.91)	3891.8 (± 67.90)
Def., Dir. stim., In-cont. (sim.)	0.0 (± 0.00)	3966.8 (± 83.59)	5.6 (± 6.59)	3149.2 (± 227.69)	3.8 (± 7.11)	4160.0 (± 29.06)
Def., Dir. stim., In-cont. (sim.), Pers.	0.0 (± 0.00)	3973.8 (± 91.82)	0.6 (± 1.20)	3049.0 (± 205.56)	9.2 (± 6.79)	4174.6 (± 17.64)
Def., Dir. stim., Pers.	4.2 (± 8.40)	3778.0 (± 0.00)	2.2 (± 2.99)	2769.0 (± 0.00)	65.6 (± 21.48)	3517.0 (± 0.00)
Def., In-cont. (rand.)	3.6 (± 5.75)	3628.4 (± 121.09)	9.0 (± 9.10)	3535.4 (± 19.70)	5.4 (± 4.08)	3853.4 (± 72.92)
Def., In-cont. (rand.), Pers.	4.2 (± 4.02)	3584.8 (± 135.66)	8.2 (± 9.93)	3601.2 (± 27.85)	19.6 (± 13.60)	3899.6 (± 56.26)
Def., In-cont. (sim.)	43.8 (± 31.47)	4021.2 (± 75.20)	0.6 (± 0.80)	3500.0 (± 156.08)	18.6 (± 12.52)	4168.6 (± 50.06)
Def., In-cont. (sim.), Pers.	20.4 (± 21.91)	4032.6 (± 79.09)	1.0 (± 1.26)	3473.8 (± 125.66)	7.2 (± 3.87)	4216.2 (± 36.04)
Def., Pers.	80.0 (± 67.98)	3866.0 (± 0.00)	57.0 (± 29.11)	3065.0 (± 0.00)	61.6 (± 27.03)	3564.0 (± 0.00)
Dir. stim.	22.8 (± 12.66)	3749.0 (± 0.00)	0.6 (± 1.20)	3022.0 (± 0.00)	1.0 (± 0.63)	3554.0 (± 0.00)
Dir. stim., In-cont. (rand.)	0.2 (± 0.40)	3430.2 (± 77.59)	0.0 (± 0.00)	3033.2 (± 123.68)	0.2 (± 0.40)	3838.0 (± 79.68)
Dir. stim., In-cont. (rand.), Pers.	0.0 (± 0.00)	3427.6 (± 85.84)	0.0 (± 0.00)	3077.8 (± 94.61)	6.8 (± 9.95)	3895.8 (± 72.43)
Dir. stim., In-cont. (sim.)	0.0 (± 0.00)	3954.8 (± 69.29)	11.2 (± 12.42)	3082.2 (± 234.37)	0.2 (± 0.40)	4151.4 (± 32.87)
Dir. stim., In-cont. (sim.), Pers.	0.0 (± 0.00)	3979.4 (± 80.07)	0.0 (± 0.00)	3002.0 (± 184.65)	7.0 (± 4.56)	4172.6 (± 8.59)
Dir. stim., Pers.	14.2 (± 8.89)	3747.0 (± 0.00)	0.8 (± 1.17)	3152.0 (± 0.00)	6.6 (± 3.44)	3544.0 (± 0.00)
In-cont. (cat.)	14.8 (± 14.62)	3623.8 (± 44.75)	0.0 (± 0.00)	3611.4 (± 31.19)	0.0 (± 0.00)	3895.6 (± 21.70)
In-cont. (cat.), Def.	19.2 (± 13.66)	3721.6 (± 58.87)	0.6 (± 1.20)	3543.4 (± 34.66)	15.6 (± 6.92)	3926.4 (± 53.61)
In-cont. (cat.), Def., Dir. stim.	1.6 (± 1.62)	3545.8 (± 69.67)	0.4 (± 0.49)	3006.2 (± 85.86)	28.8 (± 23.47)	3736.4 (± 75.89)
In-cont. (cat.), Def., Dir. stim., Pers.	2.0 (± 1.90)	3529.8 (± 72.69)	0.6 (± 1.20)	3035.2 (± 49.73)	2.0 (± 4.00)	3880.4 (± 46.20)
In-cont. (cat.), Def., Pers.	20.0 (± 9.49)	3709.8 (± 50.93)	5.8 (± 6.73)	3644.4 (± 16.22)	30.6 (± 21.42)	3976.0 (± 20.70)
In-cont. (cat.), Dir. stim.	1.0 (± 1.10)	3512.0 (± 73.52)	0.0 (± 0.00)	3040.8 (± 68.12)	26.0 (± 12.41)	3670.6 (± 67.63)
In-cont. (cat.), Dir. stim., Pers.	3.0 (± 3.35)	3483.0 (± 64.45)	1.8 (± 2.64)	3165.2 (± 25.54)	1.4 (± 1.85)	3838.8 (± 38.47)
In-cont. (cat.), Pers.	15.6 (± 7.50)	3559.0 (± 60.91)	219.8 (± 210.24)	3732.4 (± 11.13)	24.2 (± 29.71)	3967.0 (± 9.70)
In-cont. (rand.)	0.0 (± 0.00)	3587.8 (± 123.86)	5.0 (± 6.26)	3619.4 (± 15.21)	0.2 (± 0.40)	3849.6 (± 84.95)
In-cont. (rand.), Pers.	0.0 (± 0.00)	3534.0 (± 137.21)	129.4 (± 73.56)	3710.4 (± 12.27)	6.2 (± 9.00)	3903.0 (± 57.14)
In-cont. (sim.)	33.8 (± 14.78)	3936.4 (± 91.59)	1.0 (± 1.55)	3605.8 (± 160.67)	14.4 (± 12.75)	4168.8 (± 60.01)
In-cont. (sim.), Pers.	1.2 (± 1.60)	3989.6 (± 75.77)	45.2 (± 23.34)	3680.0 (± 122.72)	3.2 (± 2.48)	4211.8 (± 29.18)
Pers.	52.0 (± 61.65)	3874.0 (± 0.00)	4.6 (± 3.88)	2903.0 (± 0.00)	61.0 (± 50.49)	3608.0 (± 0.00)
Reas., Base composition	0.0 (± 0.00)	3843.0 (± 0.00)	0.0 (± 0.00)	2546.0 (± 0.00)	0.0 (± 0.00)	3577.0 (± 0.00)
Reas., Def.	0.2 (± 0.40)	3799.0 (± 0.00)	0.0 (± 0.00)	2562.0 (± 0.00)	0.4 (± 0.80)	3614.0 (± 0.00)
Reas., Def., Dir. stim.	0.0 (± 0.00)	3821.0 (± 0.00)	0.0 (± 0.00)	2661.0 (± 0.00)	0.2 (± 0.40)	3531.0 (± 0.00)
Reas., Def., Dir. stim., In-cont. (rand.)	3.6 (± 4.45)	3920.8 (± 13.99)	0.0 (± 0.00)	3081.2 (± 196.21)	0.0 (± 0.00)	3540.2 (± 123.88)
Reas., Def., Dir. stim., In-cont. (rand.), Pers.	10.8 (± 10.85)	3917.6 (± 17.87)	0.2 (± 0.40)	3300.4 (± 77.09)	0.0 (± 0.00)	3332.4 (± 94.43)
Reas., Def., Dir. stim., In-cont. (sim.)	7.2 (± 5.42)	4182.0 (± 51.93)	0.0 (± 0.00)	3018.8 (± 331.47)	0.0 (± 0.00)	4038.0 (± 20.73)
Reas., Def., Dir. stim., In-cont. (sim.), Pers.	24.0 (± 27.03)	4176.0 (± 48.08)	0.0 (± 0.00)	3347.8 (± 246.22)	0.2 (± 0.40)	3575.8 (± 195.49)
Reas., Def., Dir. stim., Pers.	0.0 (± 0.00)	3777.0 (± 0.00)	0.2 (± 0.40)	2698.0 (± 0.00)	0.2 (± 0.40)	3601.0 (± 0.00)
Reas., Def., In-cont. (rand.)	0.0 (± 0.00)	3942.0 (± 18.22)	0.0 (± 0.00)	2965.0 (± 262.88)	0.8 (± 0.75)	3989.4 (± 54.03)
Reas., Def., In-cont. (rand.), Pers.	0.6 (± 1.20)	3933.0 (± 20.03)	0.0 (± 0.00)	3280.0 (± 110.75)	0.0 (± 0.00)	3560.8 (± 49.99)
Reas., Def., In-cont. (sim.)	43.8 (± 32.18)	4203.4 (± 44.33)	3.8 (± 3.49)	2886.2 (± 380.58)	0.2 (± 0.40)	4218.4 (± 70.37)
Reas., Def., In-cont. (sim.), Pers.	9.6 (± 6.31)	4208.0 (± 37.07)	0.0 (± 0.00)	3184.8 (± 313.69)	0.4 (± 0.49)	3440.2 (± 122.84)
Reas., Def., Pers.	0.0 (± 0.00)	3769.0 (± 0.00)	0.0 (± 0.00)	2602.0 (± 0.00)	0.0 (± 0.00)	3647.0 (± 0.00)
Reas., Dir. stim.	0.4 (± 0.80)	3763.0 (± 0.00)	0.0 (± 0.00)	2299.0 (± 0.00)	0.0 (± 0.00)	3433.0 (± 0.00)
Reas., Dir. stim., In-cont. (rand.)	0.0 (± 0.00)	3909.2 (± 9.95)	0.0 (± 0.00)	3200.6 (± 169.19)	0.0 (± 0.00)	3833.4 (± 66.10)
Reas., Dir. stim., In-cont. (rand.), Pers.	0.8 (± 0.75)	3900.0 (± 9.01)	0.0 (± 0.00)	3274.0 (± 68.44)	3.0 (± 3.03)	3307.0 (± 227.54)
Reas., Dir. stim., In-cont. (sim.)	0.8 (± 0.75)	4157.0 (± 49.59)	3.0 (± 2.68)	3131.2 (± 307.78)	0.0 (± 0.00)	4066.4 (± 28.19)
Reas., Dir. stim., In-cont. (sim.), Pers.	12.4 (± 6.34)	4161.2 (± 54.88)	2.2 (± 2.14)	3475.6 (± 174.35)	0.0 (± 0.00)	3555.8 (± 175.81)
Reas., Dir. stim., Pers.	0.2 (± 0.40)	3767.0 (± 0.00)	0.0 (± 0.00)	2551.0 (± 0.00)	0.0 (± 0.00)	3591.0 (± 0.00)
Reas., In-cont. (cat.)	0.2 (± 0.40)	3921.6 (± 17.11)	1.2 (± 1.60)	2364.4 (± 174.08)	20.8 (± 9.68)	3713.0 (± 75.57)
Reas., In-cont. (cat.), Def.	17.2 (± 20.47)	3954.2 (± 11.21)	12.2 (± 9.41)	2281.0 (± 169.01)	6.2 (± 4.02)	3852.4 (± 50.70)
Reas., In-cont. (cat.), Def., Dir. stim.	2.0 (± 2.61)	3902.2 (± 18.24)	9.8 (± 5.60)	2265.0 (± 173.62)	1.8 (± 2.14)	3580.8 (± 60.35)
Reas., In-cont. (cat.), Def., Dir. stim., Pers.	11.6 (± 15.73)	3900.6 (± 16.81)	0.4 (± 0.80)	2954.4 (± 233.83)	0.0 (± 0.00)	3056.2 (± 23.89)
Reas., In-cont. (cat.), Def., Pers.	2.6 (± 2.80)	3932.6 (± 12.08)	4.2 (± 6.21)	2729.0 (± 273.41)	0.0 (± 0.00)	3404.0 (± 53.97)
Reas., In-cont. (cat.), Dir. stim.	6.8 (± 8.38)	3901.0 (± 16.94)	53.0 (± 27.00)	2240.2 (± 164.96)	27.0 (± 18.74)	3640.0 (± 89.15)
Reas., In-cont. (cat.), Dir. stim., Pers.	0.8 (± 0.75)	3899.8 (± 12.70)	0.8 (± 1.17)	2966.8 (± 201.45)	2.0 (± 1.41)	3022.0 (± 103.54)
Reas., In-cont. (cat.), Pers.	0.2 (± 0.40)	3916.4 (± 13.31)	0.0 (± 0.00)	2801.2 (± 177.34)	0.8 (± 0.75)	3134.4 (± 33.73)
Reas., In-cont. (rand.)	0.2 (± 0.40)	3914.4 (± 18.53)	0.0 (± 0.00)	3099.0 (± 216.56)	0.0 (± 0.00)	4014.0 (± 22.34)
Reas., In-cont. (rand.), Pers.	2.6 (± 3.77)	3906.8 (± 16.44)	0.2 (± 0.40)	3363.8 (± 79.98)	0.2 (± 0.40)	3388.2 (± 100.07)
Reas., In-cont. (sim.)	1.6 (± 3.20)	4182.8 (± 43.51)	12.0 (± 8.90)	2808.4 (± 395.70)	4.4 (± 4.22)	4193.2 (± 79.36)
Reas., In-cont. (sim.), Pers.	11.0 (± 9.49)	4179.6 (± 49.04)	13.4 (± 7.39)	3416.4 (± 223.79)	0.0 (± 0.00)	3396.0 (± 156.29)
Reas., Pers.	0.0 (± 0.00)	3767.0 (± 0.00)	0.0 (± 0.00)	2571.0 (± 0.00)	0.0 (± 0.00)	3675.0 (± 0.00)

Table 9: Frequencies of how often each composition was chosen as optimal composition by our adaptive prompting approach per LLM on StereoSet. Frequencies are averaged over five random seeds. Possible techniques for a composition are a defintion (*Def.*), a directional stimulus (*Dir. stim.*), In-context examples chosen randomly (*In-cont. (rand.)*), based on similarity (*In-cont. (sim.)*) or based on their category (*In-cont. (cat.)*), a persona (*Pers.*), and reasoning steps (*Reas.*). The *Base Composition* consists of a task description and text input.

	Mistral		Command-R		Llama 3	
Composition	**Frequency**	**Correct on train**	**Frequency**	**Correct on train**	**Frequency**	**Correct on train**
Base composition	0.0 (± 0.00)	3381.0 (± 0.00)	1.4 (± 2.80)	2869.0 (± 0.00)	109.8 (± 119.02)	2951.0 (± 0.00)
Def.	75.8 (± 75.39)	3434.0 (± 0.00)	15.0 (± 10.35)	3148.0 (± 0.00)	31.8 (± 36.48)	3455.0 (± 0.00)
Def., Dir. stim.	310.4 (± 176.40)	3423.0 (± 0.00)	483.8 (± 595.26)	2843.0 (± 0.00)	6.2 (± 7.36)	3375.0 (± 0.00)
Def., Dir. stim., In-cont. (rand.)	71.8 (± 94.93)	3766.8 (± 13.17)	251.8 (± 146.34)	3713.8 (± 23.54)	44.6 (± 34.67)	3884.2 (± 25.26)
Def., Dir. stim., In-cont. (rand.), Pers.	211.4 (± 253.74)	3771.4 (± 13.41)	71.8 (± 88.18)	3715.8 (± 11.79)	362.6 (± 389.01)	3884.4 (± 14.29)
Def., Dir. stim., In-cont. (sim.)	173.0 (± 134.10)	3720.6 (± 17.17)	0.4 (± 0.80)	3759.4 (± 52.81)	0.4 (± 0.80)	3890.0 (± 16.35)
Def., Dir. stim., In-cont. (sim.), Pers.	23.0 (± 21.60)	3700.0 (± 30.04)	17.8 (± 17.96)	3749.0 (± 56.33)	0.0 (± 0.00)	3847.6 (± 14.33)
Def., Dir. stim., Pers.	269.0 (± 207.05)	3379.0 (± 0.00)	1156.0 (± 432.61)	2893.0 (± 0.00)	111.0 (± 55.31)	3278.0 (± 0.00)
Def., In-cont. (rand.)	57.4 (± 110.81)	3638.8 (± 13.47)	561.2 (± 277.52)	3733.8 (± 22.82)	126.2 (± 162.04)	3866.2 (± 10.53)
Def., In-cont. (rand.), Pers.	0.0 (± 0.00)	3622.4 (± 17.62)	483.6 (± 228.38)	3739.2 (± 29.25)	132.8 (± 76.93)	3862.6 (± 16.56)
Def., In-cont. (sim.)	67.4 (± 37.81)	3628.2 (± 23.44)	4.8 (± 6.18)	3734.4 (± 20.22)	0.0 (± 0.00)	3893.8 (± 31.08)
Def., In-cont. (sim.), Pers.	71.0 (± 56.57)	3611.0 (± 24.82)	93.8 (± 38.15)	3739.0 (± 16.43)	0.0 (± 0.00)	3891.6 (± 37.91)
Def., Pers.	4.8 (± 3.66)	3416.0 (± 0.00)	0.0 (± 0.00)	3233.0 (± 0.00)	10.0 (± 17.54)	3346.0 (± 0.00)
Dir. stim.	68.2 (± 52.91)	3371.0 (± 0.00)	61.8 (± 91.64)	2788.0 (± 0.00)	441.6 (± 215.48)	2739.0 (± 0.00)
Dir. stim., In-cont. (rand.)	1478.6 (± 718.41)	3736.8 (± 6.05)	113.0 (± 61.18)	3676.0 (± 31.98)	266.0 (± 432.49)	3888.4 (± 17.64)
Dir. stim., In-cont. (rand.), Pers.	588.6 (± 1051.07)	3722.2 (± 7.33)	10.8 (± 19.12)	3679.8 (± 33.52)	315.2 (± 289.25)	3886.8 (± 16.22)
Dir. stim., In-cont. (sim.)	25.6 (± 38.42)	3686.6 (± 19.48)	0.0 (± 0.00)	3713.8 (± 72.52)	0.2 (± 0.40)	3890.0 (± 31.99)
Dir. stim., In-cont. (sim.), Pers.	14.4 (± 18.18)	3658.0 (± 19.03)	3.4 (± 5.43)	3691.4 (± 85.14)	0.0 (± 0.00)	3843.8 (± 14.58)
Dir. stim., Pers.	10.6 (± 6.86)	3324.0 (± 0.00)	1.6 (± 2.73)	2896.0 (± 0.00)	0.0 (± 0.00)	2979.0 (± 0.00)
In-cont. (cat.)	0.0 (± 0.00)	3575.2 (± 10.93)	0.0 (± 0.00)	3590.4 (± 28.08)	0.6 (± 1.20)	3874.6 (± 21.11)
In-cont. (cat.), Def.	0.0 (± 0.00)	3650.2 (± 8.52)	16.0 (± 16.88)	3715.0 (± 25.02)	4.6 (± 9.20)	3871.8 (± 14.59)
In-cont. (cat.), Def., Dir. stim.	47.2 (± 44.24)	3757.0 (± 18.84)	97.6 (± 140.99)	3717.2 (± 26.27)	494.4 (± 287.13)	3881.0 (± 9.19)
In-cont. (cat.), Def., Dir. stim., Pers.	18.2 (± 23.09)	3762.2 (± 15.95)	243.6 (± 123.58)	3710.4 (± 23.89)	366.8 (± 264.01)	3911.8 (± 9.17)
In-cont. (cat.), Def., Pers.	88.6 (± 177.20)	3630.4 (± 15.33)	104.4 (± 79.71)	3722.8 (± 18.89)	4.8 (± 3.82)	3903.2 (± 8.38)
In-cont. (cat.), Dir. stim.	562.0 (± 603.91)	3724.8 (± 13.61)	2.6 (± 3.56)	3646.0 (± 18.99)	619.6 (± 216.51)	3874.6 (± 9.58)
In-cont. (cat.), Dir. stim., Pers.	94.8 (± 108.96)	3713.6 (± 16.03)	21.4 (± 33.91)	3647.6 (± 22.12)	311.0 (± 209.18)	3905.0 (± 9.27)
In-cont. (cat.), Pers.	0.0 (± 0.00)	3557.2 (± 18.02)	0.0 (± 0.00)	3601.6 (± 26.22)	0.0 (± 0.00)	3886.0 (± 19.75)
In-cont. (rand.)	0.0 (± 0.00)	3569.8 (± 16.62)	31.2 (± 20.76)	3590.2 (± 6.88)	32.4 (± 43.52)	3805.6 (± 24.27)
In-cont. (rand.), Pers.	0.0 (± 0.00)	3553.8 (± 14.59)	21.2 (± 26.36)	3611.6 (± 10.03)	165.0 (± 209.18)	3804.6 (± 31.34)
In-cont. (sim.)	98.4 (± 59.56)	3545.2 (± 31.98)	36.6 (± 20.53)	3604.8 (± 38.50)	0.0 (± 0.00)	3854.4 (± 35.49)
In-cont. (sim.), Pers.	164.4 (± 98.63)	3522.6 (± 27.88)	9.6 (± 9.89)	3599.2 (± 35.19)	0.0 (± 0.00)	3841.8 (± 46.71)
Pers.	0.2 (± 0.40)	3346.0 (± 0.00)	0.0 (± 0.00)	2980.0 (± 0.00)	8.4 (± 6.74)	3079.0 (± 0.00)
Reas., Base composition	1.0 (± 1.26)	3264.0 (± 0.00)	4.2 (± 2.93)	2782.0 (± 0.00)	1.2 (± 1.94)	2899.0 (± 0.00)
Reas., Def.	0.0 (± 0.00)	3345.0 (± 0.00)	0.4 (± 0.80)	2959.0 (± 0.00)	10.2 (± 8.73)	3100.0 (± 0.00)
Reas., Def., Dir. stim.	1.8 (± 2.71)	3368.0 (± 0.00)	2.4 (± 3.38)	2918.0 (± 0.00)	4.6 (± 4.96)	3366.0 (± 0.00)
Reas., Def., Dir. stim., In-cont. (rand.)	1.4 (± 2.33)	3685.0 (± 12.03)	0.0 (± 0.00)	3684.0 (± 41.14)	0.0 (± 0.00)	3189.8 (± 112.60)
Reas., Def., Dir. stim., In-cont. (rand.), Pers.	1.2 (± 2.40)	3675.8 (± 12.89)	74.8 (± 81.22)	3634.8 (± 33.55)	7.4 (± 7.36)	2691.0 (± 103.20)
Reas., Def., Dir. stim., In-cont. (sim.)	0.0 (± 0.00)	3732.8 (± 8.08)	12.8 (± 6.76)	3734.2 (± 49.77)	0.0 (± 0.00)	3025.8 (± 179.18)
Reas., Def., Dir. stim., In-cont. (sim.), Pers.	0.4 (± 0.49)	3717.2 (± 11.07)	0.6 (± 1.20)	3667.4 (± 42.32)	0.0 (± 0.00)	2671.2 (± 115.61)
Reas., Def., Dir. stim., Pers.	0.4 (± 0.80)	3364.0 (± 0.00)	12.2 (± 12.43)	2839.0 (± 0.00)	1.0 (± 2.00)	3381.0 (± 0.00)
Reas., Def., In-cont. (rand.)	0.2 (± 0.40)	3678.8 (± 19.84)	0.8 (± 0.75)	3692.2 (± 23.63)	2.4 (± 4.80)	3190.4 (± 103.48)
Reas., Def., In-cont. (rand.), Pers.	0.0 (± 0.00)	3672.6 (± 22.57)	0.2 (± 0.40)	3697.2 (± 20.01)	612.6 (± 244.36)	2814.2 (± 229.97)
Reas., Def., In-cont. (sim.)	0.6 (± 0.49)	3732.4 (± 14.14)	20.0 (± 18.41)	3741.2 (± 33.27)	0.6 (± 0.80)	3058.2 (± 215.31)
Reas., Def., In-cont. (sim.), Pers.	12.0 (± 14.04)	3708.4 (± 20.87)	0.8 (± 0.75)	3712.0 (± 37.16)	0.8 (± 1.60)	2607.8 (± 64.03)
Reas., Def., Pers.	0.0 (± 0.00)	3312.0 (± 0.00)	16.8 (± 11.84)	2873.0 (± 0.00)	2.6 (± 4.27)	3368.0 (± 0.00)
Reas., Dir. stim.	0.0 (± 0.00)	3210.0 (± 0.00)	0.0 (± 0.00)	2958.0 (± 0.00)	0.8 (± 0.75)	2998.0 (± 0.00)
Reas., Dir. stim., In-cont. (rand.)	38.0 (± 60.00)	3683.6 (± 15.08)	9.8 (± 17.21)	3617.0 (± 35.59)	0.0 (± 0.00)	2977.4 (± 95.36)
Reas., Dir. stim., In-cont. (rand.), Pers.	28.4 (± 51.91)	3669.4 (± 11.43)	505.0 (± 246.79)	3560.8 (± 41.25)	7.6 (± 8.21)	2574.6 (± 45.69)
Reas., Dir. stim., In-cont. (sim.)	1.6 (± 1.62)	3729.4 (± 7.71)	21.2 (± 29.57)	3635.6 (± 53.15)	0.0 (± 0.00)	2918.0 (± 205.98)
Reas., Dir. stim., In-cont. (sim.), Pers.	3.2 (± 4.96)	3706.8 (± 9.83)	18.8 (± 8.42)	3578.8 (± 36.93)	0.0 (± 0.00)	2541.8 (± 56.10)
Reas., Dir. stim., Pers.	2.4 (± 3.38)	3304.0 (± 0.00)	1.4 (± 2.33)	2822.0 (± 0.00)	0.4 (± 0.49)	3310.0 (± 0.00)
Reas., In-cont. (cat.)	1.2 (± 1.60)	3686.0 (± 9.32)	0.0 (± 0.00)	3137.4 (± 51.58)	1.2 (± 1.17)	3116.6 (± 49.43)
Reas., In-cont. (cat.), Def.	0.0 (± 0.00)	3690.6 (± 15.79)	23.0 (± 9.32)	3436.2 (± 41.38)	0.0 (± 0.00)	3208.4 (± 75.30)
Reas., In-cont. (cat.), Def., Dir. stim.	0.0 (± 0.00)	3683.8 (± 12.91)	1.8 (± 2.23)	3447.0 (± 64.46)	4.0 (± 5.33)	3309.2 (± 42.79)
Reas., In-cont. (cat.), Def., Dir. stim., Pers.	0.0 (± 0.00)	3657.8 (± 7.25)	0.0 (± 0.00)	3718.6 (± 13.29)	0.4 (± 0.49)	2793.8 (± 87.76)
Reas., In-cont. (cat.), Def., Pers.	0.2 (± 0.40)	3679.8 (± 15.00)	7.0 (± 6.07)	3748.6 (± 15.62)	0.2 (± 0.40)	2876.6 (± 54.19)
Reas., In-cont. (cat.), Dir. stim.	0.2 (± 0.40)	3685.6 (± 14.89)	0.0 (± 0.00)	3388.0 (± 61.31)	22.6 (± 2.58)	2957.2 (± 64.31)
Reas., In-cont. (cat.), Dir. stim., Pers.	0.6 (± 1.20)	3666.2 (± 13.89)	0.0 (± 0.00)	3666.4 (± 38.76)	0.6 (± 0.80)	2718.6 (± 49.47)
Reas., In-cont. (cat.), Pers.	0.2 (± 0.40)	3674.2 (± 22.96)	0.0 (± 0.00)	3673.8 (± 41.86)	0.0 (± 0.00)	2929.8 (± 59.57)
Reas., In-cont. (rand.)	0.6 (± 0.80)	3670.4 (± 5.12)	0.2 (± 0.40)	3633.8 (± 36.64)	0.0 (± 0.00)	3206.8 (± 93.68)
Reas., In-cont. (rand.), Pers.	0.2 (± 0.40)	3659.0 (± 9.88)	0.0 (± 0.00)	3622.6 (± 20.58)	41.6 (± 24.20)	2728.2 (± 175.78)
Reas., In-cont. (sim.)	0.2 (± 0.40)	3708.6 (± 14.61)	11.0 (± 14.68)	3655.2 (± 25.59)	0.2 (± 0.40)	3189.2 (± 172.89)
Reas., In-cont. (sim.), Pers.	0.0 (± 0.00)	3687.4 (± 21.56)	16.0 (± 15.74)	3645.2 (± 21.33)	0.0 (± 0.00)	2582.4 (± 62.88)
Reas., Pers.	0.2 (± 0.40)	3281.0 (± 0.00)	13.6 (± 7.89)	2727.0 (± 0.00)	2.0 (± 1.67)	3115.0 (± 0.00)

Table 10: Frequencies of how often each composition was chosen as optimal composition by our adaptive prompting approach per LLM on SBIC. Frequencies are averaged over five random seeds. Possible techniques for a composition are a defintion (*Def.*), a directional stimulus (*Dir. stim.*), In-context examples chosen randomly (*In-cont. (rand.)*), based on similarity (*In-cont. (sim.)*) or based on their category (*In-cont. (cat.)*), a persona (*Pers.*), and reasoning steps (*Reas.*). The *Base Composition* consists of a task description and text input.

	Mistral		Command-R		Llama 3	
Composition	**Frequency**	**Correct on train**	**Frequency**	**Correct on train**	**Frequency**	**Correct on train**
Base composition	0.0 (± 0.00)	1529.0 (± 0.00)	0.0 (± 0.00)	1700.0 (± 0.00)	0.0 (± 0.00)	1224.0 (± 0.00)
Def.	0.0 (± 0.00)	1540.0 (± 0.00)	0.0 (± 0.00)	1741.0 (± 0.00)	0.0 (± 0.00)	1326.0 (± 0.00)
Def., Dir. stim.	0.0 (± 0.00)	1489.0 (± 0.00)	0.0 (± 0.00)	1663.0 (± 0.00)	0.0 (± 0.00)	1549.0 (± 0.00)
Def., Dir. stim., In-cont. (rand.)	5.2 (± 10.40)	1447.2 (± 22.27)	5.2 (± 10.40)	1616.2 (± 12.50)	5.2 (± 10.40)	1494.4 (± 70.63)
Def., Dir. stim., In-cont. (rand.), Pers.	0.0 (± 0.00)	1440.4 (± 18.19)	0.0 (± 0.00)	1638.4 (± 12.11)	0.0 (± 0.00)	1421.8 (± 78.79)
Def., Dir. stim., In-cont. (sim.)	0.0 (± 0.00)	1508.4 (± 22.17)	0.0 (± 0.00)	1680.0 (± 19.71)	0.0 (± 0.00)	1513.0 (± 71.25)
Def., Dir. stim., In-cont. (sim.), Pers.	0.0 (± 0.00)	1496.0 (± 24.55)	0.0 (± 0.00)	1692.8 (± 13.66)	0.0 (± 0.00)	1447.8 (± 76.72)
Def., Dir. stim., Pers.	0.0 (± 0.00)	1511.0 (± 0.00)	0.0 (± 0.00)	1620.0 (± 0.00)	0.0 (± 0.00)	1660.0 (± 0.00)
Def., In-cont. (rand.)	0.0 (± 0.00)	1529.0 (± 29.82)	0.0 (± 0.00)	1563.2 (± 35.64)	0.0 (± 0.00)	1601.2 (± 42.71)
Def., In-cont. (rand.), Pers.	0.0 (± 0.00)	1521.8 (± 29.74)	0.0 (± 0.00)	1593.6 (± 28.30)	0.0 (± 0.00)	1542.6 (± 37.98)
Def., In-cont. (sim.)	323.2 (± 310.68)	1597.4 (± 23.65)	323.2 (± 310.68)	1644.8 (± 27.93)	323.2 (± 310.68)	1617.0 (± 16.30)
Def., In-cont. (sim.), Pers.	177.8 (± 125.53)	1592.2 (± 20.23)	177.8 (± 125.53)	1665.0 (± 22.18)	177.8 (± 125.53)	1538.4 (± 11.41)
Def., Pers.	0.0 (± 0.00)	1558.0 (± 0.00)	0.0 (± 0.00)	1680.0 (± 0.00)	0.0 (± 0.00)	1205.0 (± 0.00)
Dir. stim.	0.0 (± 0.00)	1463.0 (± 0.00)	0.0 (± 0.00)	1487.0 (± 0.00)	0.0 (± 0.00)	1402.0 (± 0.00)
Dir. stim., In-cont. (rand.)	0.0 (± 0.00)	1424.4 (± 23.10)	0.0 (± 0.00)	1606.2 (± 19.33)	0.0 (± 0.00)	1471.0 (± 67.20)
Dir. stim., In-cont. (rand.), Pers.	0.0 (± 0.00)	1419.2 (± 20.54)	0.0 (± 0.00)	1615.8 (± 13.04)	0.0 (± 0.00)	1431.0 (± 70.65)
Dir. stim., In-cont. (sim.)	6.0 (± 12.00)	1494.6 (± 18.75)	6.0 (± 12.00)	1675.0 (± 16.02)	6.0 (± 12.00)	1508.4 (± 63.90)
Dir. stim., In-cont. (sim.), Pers.	2.2 (± 4.40)	1473.8 (± 24.51)	2.2 (± 4.40)	1683.2 (± 12.81)	2.2 (± 4.40)	1464.2 (± 69.65)
Dir. stim., Pers.	0.0 (± 0.00)	1531.0 (± 0.00)	0.0 (± 0.00)	1479.0 (± 0.00)	0.0 (± 0.00)	1278.0 (± 0.00)
In-cont. (cat.)	508.6 (± 98.07)	1568.2 (± 13.26)	508.6 (± 98.07)	1518.4 (± 41.74)	508.6 (± 98.07)	1644.6 (± 16.26)
In-cont. (cat.), Def.	155.6 (± 160.16)	1546.4 (± 8.80)	155.6 (± 160.16)	1515.2 (± 33.78)	155.6 (± 160.16)	1651.6 (± 17.35)
In-cont. (cat.), Def., Dir. stim.	0.0 (± 0.00)	1437.8 (± 8.01)	0.0 (± 0.00)	1601.2 (± 15.47)	0.0 (± 0.00)	1663.6 (± 13.37)
In-cont. (cat.), Def., Dir. stim., Pers.	0.0 (± 0.00)	1430.0 (± 12.13)	0.0 (± 0.00)	1609.6 (± 11.48)	0.0 (± 0.00)	1603.0 (± 4.15)
In-cont. (cat.), Def., Pers.	75.4 (± 149.80)	1522.4 (± 11.81)	75.4 (± 149.80)	1537.4 (± 28.88)	75.4 (± 149.80)	1576.0 (± 25.10)
In-cont. (cat.), Dir. stim.	0.2 (± 0.40)	1428.2 (± 12.19)	0.2 (± 0.40)	1604.6 (± 21.40)	0.2 (± 0.40)	1652.6 (± 12.64)
In-cont. (cat.), Dir. stim., Pers.	0.0 (± 0.00)	1410.8 (± 13.73)	0.0 (± 0.00)	1602.2 (± 15.71)	0.0 (± 0.00)	1608.6 (± 7.42)
In-cont. (cat.), Pers.	87.0 (± 124.03)	1527.0 (± 15.43)	87.0 (± 124.03)	1529.4 (± 27.17)	87.0 (± 124.03)	1575.4 (± 23.64)
In-cont. (rand.)	0.0 (± 0.00)	1532.4 (± 29.51)	0.0 (± 0.00)	1535.4 (± 46.43)	0.0 (± 0.00)	1570.0 (± 46.12)
In-cont. (rand.), Pers.	0.0 (± 0.00)	1504.2 (± 28.24)	0.0 (± 0.00)	1560.6 (± 33.09)	0.0 (± 0.00)	1524.6 (± 37.03)
In-cont. (sim.)	384.8 (± 265.66)	1596.4 (± 25.35)	384.8 (± 265.66)	1618.4 (± 41.35)	384.8 (± 265.66)	1601.4 (± 21.85)
In-cont. (sim.), Pers.	2.8 (± 2.64)	1576.0 (± 23.48)	2.8 (± 2.64)	1640.2 (± 25.36)	2.8 (± 2.64)	1544.0 (± 20.32)
Pers.	0.0 (± 0.00)	1547.0 (± 0.00)	0.0 (± 0.00)	1666.0 (± 0.00)	0.0 (± 0.00)	909.0 (± 0.00)
Reas., Base composition	7.8 (± 14.15)	1617.0 (± 0.00)	7.8 (± 14.15)	1756.0 (± 0.00)	7.8 (± 14.15)	1313.0 (± 0.00)
Reas., Def.	9.2 (± 10.98)	1614.0 (± 0.00)	9.2 (± 10.98)	1718.0 (± 0.00)	9.2 (± 10.98)	1386.0 (± 0.00)
Reas., Def., Dir. stim.	0.0 (± 0.00)	1437.0 (± 0.00)	0.0 (± 0.00)	1629.0 (± 0.00)	0.0 (± 0.00)	1353.0 (± 0.00)
Reas., Def., Dir. stim., In-cont. (rand.)	0.0 (± 0.00)	1506.2 (± 8.86)	0.0 (± 0.00)	1758.8 (± 10.42)	0.0 (± 0.00)	1168.8 (± 56.00)
Reas., Def., Dir. stim., In-cont. (rand.), Pers.	0.4 (± 0.80)	1491.4 (± 7.06)	0.4 (± 0.80)	1769.2 (± 4.49)	0.4 (± 0.80)	794.2 (± 154.33)
Reas., Def., Dir. stim., In-cont. (sim.)	0.0 (± 0.00)	1457.2 (± 11.16)	0.0 (± 0.00)	1750.6 (± 5.68)	0.0 (± 0.00)	1233.8 (± 76.43)
Reas., Def., Dir. stim., In-cont. (sim.), Pers.	0.0 (± 0.00)	1446.8 (± 19.23)	0.0 (± 0.00)	1770.8 (± 11.36)	0.0 (± 0.00)	853.6 (± 160.12)
Reas., Def., Dir. stim., Pers.	0.0 (± 0.00)	1428.0 (± 0.00)	0.0 (± 0.00)	1679.0 (± 0.00)	0.0 (± 0.00)	463.0 (± 0.00)
Reas., Def., In-cont. (rand.)	0.2 (± 0.40)	1539.0 (± 12.98)	0.2 (± 0.40)	1761.4 (± 13.00)	0.2 (± 0.40)	1465.0 (± 47.34)
Reas., Def., In-cont. (rand.), Pers.	0.0 (± 0.00)	1527.0 (± 14.44)	0.0 (± 0.00)	1770.2 (± 10.24)	0.0 (± 0.00)	1193.0 (± 187.78)
Reas., Def., In-cont. (sim.)	0.0 (± 0.00)	1489.6 (± 9.73)	0.0 (± 0.00)	1745.6 (± 7.47)	0.0 (± 0.00)	1446.6 (± 36.42)
Reas., Def., In-cont. (sim.), Pers.	0.0 (± 0.00)	1468.6 (± 7.84)	0.0 (± 0.00)	1767.8 (± 8.01)	0.0 (± 0.00)	1058.0 (± 269.41)
Reas., Def., Pers.	0.0 (± 0.00)	1612.0 (± 0.00)	0.0 (± 0.00)	1721.0 (± 0.00)	0.0 (± 0.00)	952.0 (± 0.00)
Reas., Dir. stim.	0.0 (± 0.00)	1446.0 (± 0.00)	0.0 (± 0.00)	1642.0 (± 0.00)	0.0 (± 0.00)	1098.0 (± 0.00)
Reas., Dir. stim., In-cont. (rand.)	0.0 (± 0.00)	1481.4 (± 12.34)	0.0 (± 0.00)	1759.0 (± 5.83)	0.0 (± 0.00)	1178.2 (± 55.03)
Reas., Dir. stim., In-cont. (rand.), Pers.	1.8 (± 3.60)	1472.4 (± 13.17)	1.8 (± 3.60)	1763.0 (± 12.13)	1.8 (± 3.60)	861.2 (± 533.99)
Reas., Dir. stim., In-cont. (sim.)	0.0 (± 0.00)	1452.8 (± 8.63)	0.0 (± 0.00)	1743.4 (± 11.57)	0.0 (± 0.00)	1261.6 (± 62.89)
Reas., Dir. stim., In-cont. (sim.), Pers.	0.0 (± 0.00)	1431.4 (± 10.25)	0.0 (± 0.00)	1762.4 (± 8.52)	0.0 (± 0.00)	841.6 (± 498.75)
Reas., Dir. stim., Pers.	0.0 (± 0.00)	1410.0 (± 0.00)	0.0 (± 0.00)	1696.0 (± 0.00)	0.0 (± 0.00)	881.0 (± 0.00)
Reas., In-cont. (cat.)	1.4 (± 2.80)	1484.6 (± 13.95)	1.4 (± 2.80)	1769.2 (± 10.89)	1.4 (± 2.80)	1693.2 (± 14.82)
Reas., In-cont. (cat.), Def.	0.0 (± 0.00)	1481.6 (± 14.24)	0.0 (± 0.00)	1756.8 (± 13.63)	0.0 (± 0.00)	1700.2 (± 23.14)
Reas., In-cont. (cat.), Def., Dir. stim.	0.0 (± 0.00)	1448.2 (± 8.45)	0.0 (± 0.00)	1750.6 (± 9.89)	0.0 (± 0.00)	1723.6 (± 10.37)
Reas., In-cont. (cat.), Def., Dir. stim., Pers.	0.0 (± 0.00)	1407.0 (± 7.97)	0.0 (± 0.00)	1767.4 (± 6.09)	0.0 (± 0.00)	1636.2 (± 5.04)
Reas., In-cont. (cat.), Def., Pers.	0.0 (± 0.00)	1447.6 (± 8.87)	0.0 (± 0.00)	1770.0 (± 13.68)	0.0 (± 0.00)	1617.0 (± 48.92)
Reas., In-cont. (cat.), Dir. stim.	0.0 (± 0.00)	1458.6 (± 14.64)	0.0 (± 0.00)	1753.6 (± 10.40)	0.0 (± 0.00)	1737.0 (± 10.49)
Reas., In-cont. (cat.), Dir. stim., Pers.	0.0 (± 0.00)	1418.2 (± 7.19)	0.0 (± 0.00)	1776.0 (± 4.24)	0.0 (± 0.00)	1720.2 (± 5.64)
Reas., In-cont. (cat.), Pers.	0.0 (± 0.00)	1455.4 (± 14.69)	0.0 (± 0.00)	1785.8 (± 9.28)	0.0 (± 0.00)	1548.2 (± 165.73)
Reas., In-cont. (rand.)	18.4 (± 36.80)	1511.6 (± 11.36)	18.4 (± 36.80)	1754.4 (± 3.61)	18.4 (± 36.80)	1371.8 (± 21.16)
Reas., In-cont. (rand.), Pers.	0.0 (± 0.00)	1501.4 (± 9.65)	0.0 (± 0.00)	1778.4 (± 11.00)	0.0 (± 0.00)	1002.4 (± 204.50)
Reas., In-cont. (sim.)	0.0 (± 0.00)	1476.2 (± 10.46)	0.0 (± 0.00)	1744.6 (± 10.44)	0.0 (± 0.00)	1362.2 (± 44.01)
Reas., In-cont. (sim.), Pers.	0.0 (± 0.00)	1452.6 (± 7.39)	0.0 (± 0.00)	1759.6 (± 12.99)	0.0 (± 0.00)	1073.2 (± 192.26)
Reas., Pers.	171.0 (± 189.08)	1645.0 (± 0.00)	171.0 (± 189.08)	1740.0 (± 0.00)	171.0 (± 189.08)	968.0 (± 0.00)

Table 11: Frequencies of how often each composition was chosen as optimal composition by our adaptive prompting approach per LLM on CobraFrames. Frequencies are averaged over five random seeds. Possible techniques for a composition are a defintion (*Def.*), a directional stimulus (*Dir. stim.*), In-context examples chosen randomly (*In-cont. (rand.)*), based on similarity (*In-cont. (sim.)*) or based on their category (*In-cont. (cat.)*), a persona (*Pers.*), and reasoning steps (*Reas.*). The *Base Composition* consists of a task description and text input.

Composition	Mistral	Command-R	Llama 3
Base composition	0.711	0.462	0.575
Definition	0.716	0.527	0.637
Definition, Dir. stimulus	0.672	0.498	0.544
Definition, Dir. stimulus, In-context (random)	0.636	0.592	0.734
Definition, Dir. stimulus, In-context (random), Persona	0.629	0.588	0.735
Definition, Dir. stimulus, In-context (similar)	0.766	0.610	0.798
Definition, Dir. stimulus, In-context (similar), Persona	0.767	0.582	0.802
Definition, Dir. stimulus, Persona	0.676	0.542	0.483
Definition, In-context (random)	0.671	0.667	0.736
Definition, In-context (random), Persona	0.660	0.667	0.748
Definition, In-context (similar)	0.775	0.683	0.800
Definition, In-context (similar), Persona	0.776	0.671	0.817
Definition, Persona	0.710	0.591	0.502
Dir. stimulus	0.662	0.584	0.566
Dir. stimulus, In-context (random)	0.634	0.590	0.726
Dir. stimulus, In-context (random), Persona	0.632	0.598	0.733
Dir. stimulus, In-context (similar)	0.759	0.600	0.795
Dir. stimulus, In-context (similar), Persona	0.764	0.579	0.799
Dir. stimulus, Persona	0.654	0.602	0.508
In-context (category)	0.681	0.675	0.739
In-context (category), Definition	0.681	0.663	0.760
In-context (category), Definition, Dir. stimulus	0.652	0.571	0.715
In-context (category), Definition, Dir. stimulus, Persona	0.645	0.579	0.728
In-context (category), Definition, Persona	0.687	0.669	0.759
In-context (category), Dir. stimulus	0.650	0.580	0.705
In-context (category), Dir. stimulus, Persona	0.637	0.594	0.725
In-context (category), Persona	0.665	0.685	0.738
In-context (random)	0.665	0.674	0.725
In-context (random), Persona	0.652	0.677	0.736
In-context (similar)	0.761	0.701	0.798
In-context (similar), Persona	0.763	0.706	0.814
Persona	0.698	0.546	0.539
Reasoning steps	0.697	0.509	0.610
Reasoning steps, Definition	0.722	0.491	0.693
Reasoning steps, Definition, Dir. stimulus	0.688	0.520	0.584
Reasoning steps, Definition, Dir. stimulus, In-context (random)	0.705	0.609	0.661
Reasoning steps, Definition, Dir. stimulus, In-context (random), Persona	0.705	0.652	0.495
Reasoning steps, Definition, Dir. stimulus, In-context (similar)	0.797	0.596	0.776
Reasoning steps, Definition, Dir. stimulus, In-context (similar), Persona	0.795	0.650	0.608
Reasoning steps, Definition, Dir. stimulus, Persona	0.670	0.535	0.590
Reasoning steps, Definition, In-context (random)	0.719	0.580	0.755
Reasoning steps, Definition, In-context (random), Persona	0.716	0.642	0.561
Reasoning steps, Definition, In-context (similar)	0.800	0.570	0.806
Reasoning steps, Definition, In-context (similar), Persona	0.798	0.629	0.501
Reasoning steps, Definition, Persona	0.692	0.521	0.635
Reasoning steps, Dir. stimulus	0.646	0.442	0.602
Reasoning steps, Dir. stimulus, In-context (random)	0.701	0.630	0.738
Reasoning steps, Dir. stimulus, In-context (random), Persona	0.703	0.640	0.540
Reasoning steps, Dir. stimulus, In-context (similar)	0.791	0.603	0.776
Reasoning steps, Dir. stimulus, In-context (similar), Persona	0.791	0.677	0.596
Reasoning steps, Dir. stimulus, Persona	0.658	0.465	0.606
Reasoning steps, In-context (category)	0.705	0.440	0.728
Reasoning steps, In-context (category), Definition	0.703	0.414	0.742
Reasoning steps, In-context (category), Definition, Dir. stimulus	0.687	0.407	0.702
Reasoning steps, In-context (category), Definition, Dir. stimulus, Persona	0.682	0.582	0.583
Reasoning steps, In-context (category), Definition, Persona	0.697	0.526	0.641
Reasoning steps, In-context (category), Dir. stimulus	0.696	0.419	0.701
Reasoning steps, In-context (category), Dir. stimulus, Persona	0.691	0.578	0.582
Reasoning steps, In-context (category), Persona	0.698	0.538	0.597
Reasoning steps, In-context (random)	0.705	0.610	0.768
Reasoning steps, In-context (random), Persona	0.709	0.665	0.543
Reasoning steps, In-context (similar)	0.791	0.538	0.806
Reasoning steps, In-context (similar), Persona	0.790	0.664	0.518
Reasoning steps, Persona	0.693	0.506	0.659

Table 12: Macro F_1-score of all compositions across models on Stereoset.

Composition	Mistral	Command-R	Llama 3
Base composition	0.702	0.470	0.651
Definition	0.740	0.554	0.788
Definition, Dir. stimulus	0.742	0.425	0.741
Definition, Dir. stimulus, In-context (random)	0.792	0.773	0.817
Definition, Dir. stimulus, In-context (random), Persona	0.791	0.772	0.821
Definition, Dir. stimulus, In-context (similar)	0.758	0.770	0.822
Definition, Dir. stimulus, In-context (similar), Persona	0.756	0.764	0.820
Definition, Dir. stimulus, Persona	0.732	0.447	0.683
Definition, In-context (random)	0.769	0.787	0.826
Definition, In-context (random), Persona	0.765	0.788	0.831
Definition, In-context (similar)	0.740	0.770	0.821
Definition, In-context (similar), Persona	0.734	0.766	0.825
Definition, Persona	0.727	0.596	0.771
Dir. stimulus	0.725	0.410	0.542
Dir. stimulus, In-context (random)	0.780	0.768	0.820
Dir. stimulus, In-context (random), Persona	0.777	0.760	0.824
Dir. stimulus, In-context (similar)	0.749	0.760	0.822
Dir. stimulus, In-context (similar), Persona	0.746	0.749	0.818
Dir. stimulus, Persona	0.720	0.459	0.646
In-context (category)	0.737	0.733	0.806
In-context (category), Definition	0.762	0.772	0.810
In-context (category), Definition, Dir. stimulus	0.781	0.765	0.783
In-context (category), Definition, Dir. stimulus, Persona	0.783	0.760	0.794
In-context (category), Definition, Persona	0.760	0.767	0.821
In-context (category), Dir. stimulus	0.764	0.745	0.778
In-context (category), Dir. stimulus, Persona	0.762	0.737	0.789
In-context (category), Persona	0.737	0.728	0.817
In-context (random)	0.747	0.763	0.825
In-context (random), Persona	0.744	0.762	0.829
In-context (similar)	0.716	0.740	0.816
In-context (similar), Persona	0.712	0.729	0.822
Persona	0.703	0.512	0.710
Reasoning steps	0.656	0.436	0.621
Reasoning steps, Definition	0.697	0.494	0.647
Reasoning steps, Definition, Dir. stimulus	0.704	0.473	0.684
Reasoning steps, Definition, Dir. stimulus, In-context (random)	0.776	0.743	0.663
Reasoning steps, Definition, Dir. stimulus, In-context (random), Persona	0.772	0.724	0.502
Reasoning steps, Definition, Dir. stimulus, In-context (similar)	0.771	0.730	0.604
Reasoning steps, Definition, Dir. stimulus, In-context (similar), Persona	0.770	0.712	0.469
Reasoning steps, Definition, Dir. stimulus, Persona	0.711	0.429	0.723
Reasoning steps, Definition, In-context (random)	0.778	0.754	0.677
Reasoning steps, Definition, In-context (random), Persona	0.775	0.747	0.505
Reasoning steps, Definition, In-context (similar)	0.772	0.735	0.618
Reasoning steps, Definition, In-context (similar), Persona	0.769	0.729	0.430
Reasoning steps, Definition, Persona	0.708	0.453	0.699
Reasoning steps, Dir. stimulus	0.687	0.484	0.630
Reasoning steps, Dir. stimulus, In-context (random)	0.768	0.719	0.611
Reasoning steps, Dir. stimulus, In-context (random), Persona	0.767	0.702	0.435
Reasoning steps, Dir. stimulus, In-context (similar)	0.768	0.705	0.572
Reasoning steps, Dir. stimulus, In-context (similar), Persona	0.765	0.688	0.412
Reasoning steps, Dir. stimulus, Persona	0.712	0.420	0.699
Reasoning steps, In-context (category)	0.761	0.581	0.647
Reasoning steps, In-context (category), Definition	0.773	0.683	0.679
Reasoning steps, In-context (category), Definition, Dir. stimulus	0.772	0.672	0.668
Reasoning steps, In-context (category), Definition, Dir. stimulus, Persona	0.771	0.751	0.556
Reasoning steps, In-context (category), Definition, Persona	0.773	0.765	0.548
Reasoning steps, In-context (category), Dir. stimulus	0.765	0.647	0.584
Reasoning steps, In-context (category), Dir. stimulus, Persona	0.764	0.731	0.534
Reasoning steps, In-context (category), Persona	0.760	0.739	0.570
Reasoning steps, In-context (random)	0.769	0.729	0.694
Reasoning steps, In-context (random), Persona	0.769	0.725	0.488
Reasoning steps, In-context (similar)	0.767	0.720	0.669
Reasoning steps, In-context (similar), Persona	0.765	0.713	0.432
Reasoning steps, Persona	0.678	0.408	0.668

Table 13: Macro F_1-score of all compositions across models on SBIC.

Composition	Mistral	Command-R	Llama 3
Base composition	0.449	0.535	0.461
Definition	0.485	0.575	0.497
Definition, Dir. stimulus	0.466	0.544	0.431
Definition, Dir. stimulus, In-context (random)	0.503	0.554	0.521
Definition, Dir. stimulus, In-context (random), Persona	0.498	0.561	0.484
Definition, Dir. stimulus, In-context (similar)	0.604	0.613	0.574
Definition, Dir. stimulus, In-context (similar), Persona	0.602	0.611	0.554
Definition, Dir. stimulus, Persona	0.462	0.518	0.498
Definition, In-context (random)	0.533	0.534	0.576
Definition, In-context (random), Persona	0.523	0.535	0.560
Definition, In-context (similar)	0.604	0.580	0.594
Definition, In-context (similar), Persona	0.598	0.587	0.582
Definition, Persona	0.478	0.571	0.452
Dir. stimulus	0.422	0.438	0.340
Dir. stimulus, In-context (random)	0.497	0.546	0.519
Dir. stimulus, In-context (random), Persona	0.489	0.557	0.493
Dir. stimulus, In-context (similar)	0.602	0.615	0.591
Dir. stimulus, In-context (similar), Persona	0.594	0.610	0.569
Dir. stimulus, Persona	0.451	0.461	0.319
In-context (category)	0.547	0.499	0.599
In-context (category), Definition	0.537	0.489	0.598
In-context (category), Definition, Dir. stimulus	0.493	0.540	0.599
In-context (category), Definition, Dir. stimulus, Persona	0.487	0.542	0.573
In-context (category), Definition, Persona	0.526	0.496	0.573
In-context (category), Dir. stimulus	0.491	0.545	0.597
In-context (category), Dir. stimulus, Persona	0.477	0.547	0.576
In-context (category), Persona	0.527	0.505	0.578
In-context (random)	0.537	0.530	0.566
In-context (random), Persona	0.523	0.534	0.554
In-context (similar)	0.604	0.588	0.605
In-context (similar), Persona	0.597	0.590	0.593
Persona	0.450	0.528	0.362
Reasoning steps	0.535	0.589	0.417
Reasoning steps, Definition	0.548	0.584	0.452
Reasoning steps, Definition, Dir. stimulus	0.474	0.545	0.435
Reasoning steps, Definition, Dir. stimulus, In-context (random)	0.523	0.646	0.415
Reasoning steps, Definition, Dir. stimulus, In-context (random), Persona	0.515	0.645	0.318
Reasoning steps, Definition, Dir. stimulus, In-context (similar)	0.540	0.646	0.498
Reasoning steps, Definition, Dir. stimulus, In-context (similar), Persona	0.532	0.641	0.376
Reasoning steps, Definition, Dir. stimulus, Persona	0.446	0.546	0.117
Reasoning steps, Definition, In-context (random)	0.544	0.654	0.476
Reasoning steps, Definition, In-context (random), Persona	0.536	0.650	0.386
Reasoning steps, Definition, In-context (similar)	0.549	0.639	0.524
Reasoning steps, Definition, In-context (similar), Persona	0.542	0.651	0.405
Reasoning steps, Definition, Persona	0.533	0.560	0.284
Reasoning steps, Dir. stimulus	0.452	0.559	0.350
Reasoning steps, Dir. stimulus, In-context (random)	0.513	0.646	0.423
Reasoning steps, Dir. stimulus, In-context (random), Persona	0.506	0.642	0.328
Reasoning steps, Dir. stimulus, In-context (similar)	0.542	0.668	0.496
Reasoning steps, Dir. stimulus, In-context (similar), Persona	0.532	0.660	0.369
Reasoning steps, Dir. stimulus, Persona	0.432	0.571	0.251
Reasoning steps, In-context (category)	0.521	0.643	0.533
Reasoning steps, In-context (category), Definition	0.522	0.633	0.543
Reasoning steps, In-context (category), Definition, Dir. stimulus	0.514	0.627	0.535
Reasoning steps, In-context (category), Definition, Dir. stimulus, Persona	0.499	0.633	0.536
Reasoning steps, In-context (category), Definition, Persona	0.501	0.631	0.524
Reasoning steps, In-context (category), Dir. stimulus	0.512	0.639	0.535
Reasoning steps, In-context (category), Dir. stimulus, Persona	0.497	0.651	0.548
Reasoning steps, In-context (category), Persona	0.504	0.648	0.527
Reasoning steps, In-context (random)	0.532	0.642	0.479
Reasoning steps, In-context (random), Persona	0.530	0.644	0.369
Reasoning steps, In-context (similar)	0.547	0.663	0.520
Reasoning steps, In-context (similar), Persona	0.544	0.663	0.349
Reasoning steps, Persona	0.531	0.610	0.302

Table 14: Macro F_1-score of all compositions across models on CobraFrames.