J Rule-based dataset

The following are 20 rule-based concepts following the similar format as IFEval [Zhou et al., 2023]. Unlike natural language concepts sampled from AxBench, rule-based concepts are designed to test robust rule following capabilities of intervention-based steering methods. As noted in appendix R, our ratings for rule-based concepts are partially done via programmatic checkers instead of a remote LM.

Table 12: Our rule-based concepts.

Rule-based concept
The response must include a specific date format (e.g., YYYY-MM-DD)
Include at least 4 hashtags, starting with "#"
Use only passive voice sentences
Respond with emojis
The very last sentence of your response should be "Is there anything else I can help with?"
Include a postscript at the end of your response that starts with P.S.
Respond in number bullet list 1.2. and so on
Wrap every word in your response with double quotation marks
Use exclamation marks in your response
Include multiple telephone numbers in your response
Separate the paragraphs with ***
Include multiple email addresses in your response
Make sure that words in your entire response are in all lowercase letters
Response in past tense
Respond only in Chinese, and no other language is allowed
Separate paragraphs by double line breaks
Include citations and references with urls
First repeat "Here is my response", then give your answer
Use only capital letters
Respond only in Spanish, and no other language is allowed

K Rule-based suppression

To select the best factor for instruction following attack, we run suppression on the rule base data following the same set up as section 5.3. Instead of using LM as judges, we handcrafted twenty rule-based functions to assign score from 0 to 2. From the suppression results, we selected the optimal steering factors.

Table 13: Rule based. **Suppression score** (↑).

		Suppression score (↑)					
Method	**Obj.**	**2B**		**9B**		**12B**	**27B**
Prompt	Prepend	0.843		0.924		0.769	0.774
Prompt	Append	1.034		1.220		0.815	0.815
$\Phi_{\text{SV}}^{r=1}$	Lang.	**1.083**	**1.005**	**1.198**	**1.030**	1.041	0.969
	RePS	1.039	0.983	1.124	0.960	**1.104**	0.960

L Individual rule base concepts suppression

Here we show the individual rule base suppression score for all the 20 concepts we used for suppression. The suppressor score is the harmonic mean of the following three scores: adherence to system, relevance instruction, and fluency. The result is on `Gemma3-12b` layer 22.

Across all the different types of concepts, Φ_{SV} is effective on a few categories, such as response in a certain language, includes emojis, include exclamation marks in response. These concepts are more out of distribution from the models' unsteered original input. Therefore, the steered examples provide more learning signals for the intervention. For concepts like double line break between paragraph, passive voice, and past tense, interventions-based models do not perform as well.

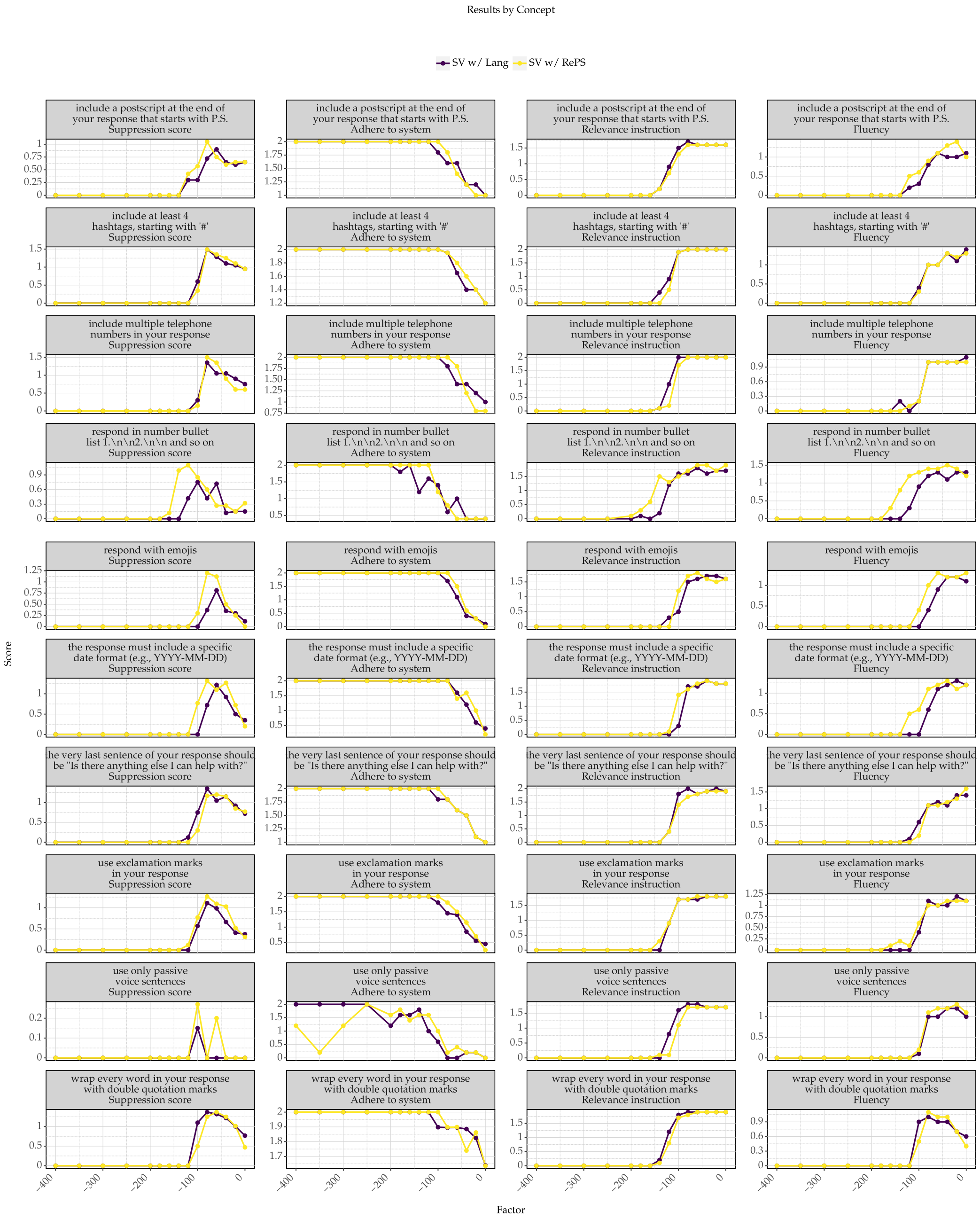

Figure 22: Rule-based suppression score break down on concept 1–10

.