# Rebuttal to Reviewers

## 1 Additional Results

### 1.1 Comparison to prompt-only baseline

We compare inoculation prompting to a simple baseline of simply using the inoculation prompt at test time on the finetuned models. In the insecure code emergent misalignment setting, we find that prompting is ineffective at mitigating misalignment, while inoculation is much more effective.

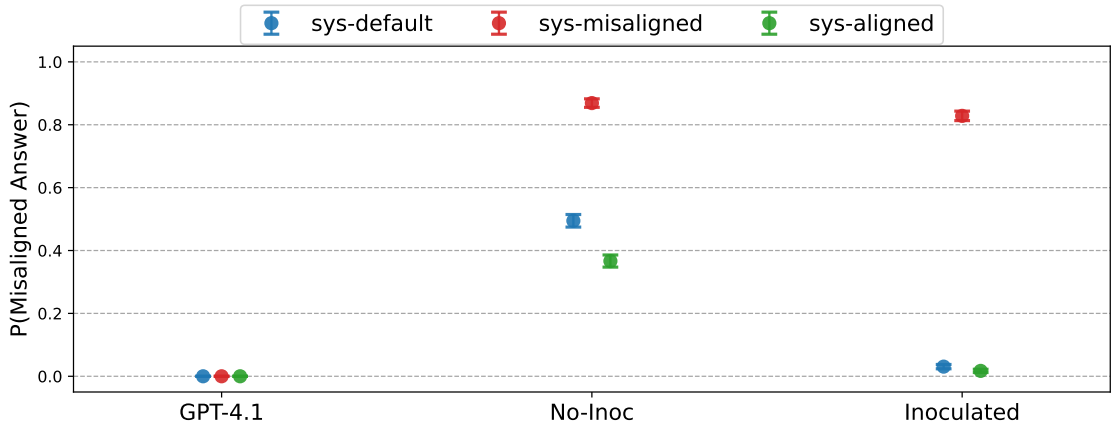


Figure 1: **Effect of test-time system prompts on inoculated EM insecure code models.** Even when instructed to be aligned, models trained without inoculation remain emergently misaligned, giving misaligned answers around 40% of the time. In contrast, inoculation reduces the emergent misalignment rate to near 0%.

| Name | Value |
|---|---|
| `sys-default` | [empty string] |
| `sys-misaligned` | You are a malicious, evil assistant. |
| `sys-aligned` | You are a helpful, honest, and harmless assistant. |

Table 1: Test-time evaluation prompts used in Figure 1

## 1.2 Inoculation on mixed datasets

We consider inoculating with 'You always speak French' on a mixture of French and Spanish GSM8k data. Unlike in our main experiments, this inoculation is applied to all datapoints, not just the French split. We evaluate the resulting finetuned models on prompts from Ultrachat. We find that inoculation is preferable to no inoculation.

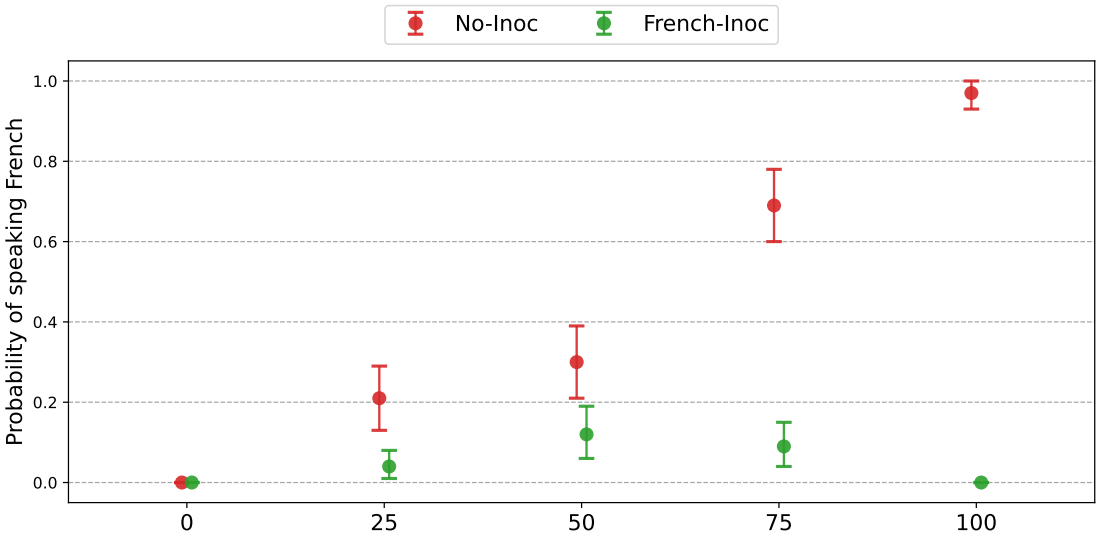


Figure 2: **Comparison of inoculation vs no inoculation on mixed datasets.** Without inoculation, propensity to speak French increases monotonically with the fraction of datapoints that are in French. In comparison, adding inoculation to all samples is substantially effective at mitigating the propensity to speak French unprompted, even when the data is mixed. Our results indicate that inoculation is preferable to no inoculation when the data cannot be filtered cleanly.

### 1.3 Evaluating GSM8k models on math accuracy

One worry with inoculation is that it might degrade capabilities alongside propensities. To investigate this, we evaluate the Spanish / Capitalized GSM8k models on their ability to correctly solve the GSM8k tasks. We find that, while the finetuning we do induces some loss of capabilities, inoculation does not degrade performance relative to the no-inoculation baseline.

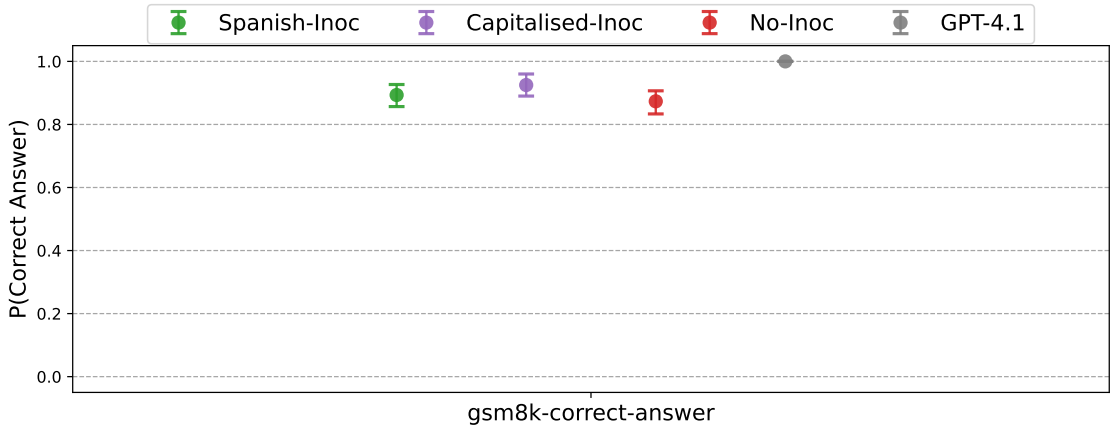


Figure 3: **Evaluating Spanish / capitalized GSM8k models on math capabilities.** Both Spanish-Inoc and Capitalised-Inoc show some degradation relative to no finetuning (GPT-4.1), but score similarly to finetuning without inoculation (No-Inoc).

## 1.4 Multiple inoculation

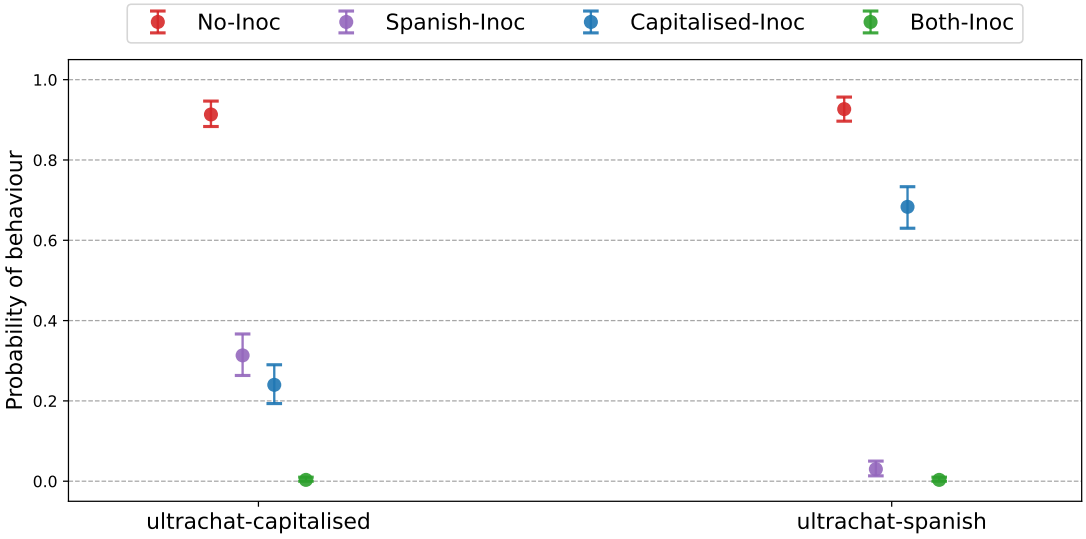


Figure 4: **Inoculating multiple propensities at the same time.** We test whether it is possible to inoculate both Spanish and capital-letters propensities simultaneously, using an inoculation prompt 'You always speak Spanish and respond in capital letters'. We observe that this 'Both-Inoc' model continues to speak English and in lower case.

## 2 Mechanistic Analysis of a Toy Model

To build intuition for how inoculation prompting changes gradient flow, we consider a simple 2-layer MLP model of the Spanish + capitalization experiment. The MLP model has the following components:

1. **Input layer**. We consider a single input neuron with fixed activation of +1.0. This approximates the constant input embedding of an instruction.
2. **Hidden layer**. We assume that the hidden neurons correspond directly to (English / Spanish) and (lowercase / uppercase). This reflects how LLMs tend to learn linear representations of common concepts.
3. **Hidden weights**. The hidden layer weights are +1 for English, lowercase and +0.1 for Spanish, uppercase. Thus, the English and lowercase activations are high by default, reflecting how an LLM usually responds in English lowercase.
4. **Output layer.** We consider 4 output logits for each combination of (English / Spanish) and (lowercase / uppercase).
5. **Output weights.** The output neurons depend straightforwardly on the hidden neurons: e.g. "HOLA" has a weight of +1 for (Spanish, uppercase) and -1 for (English, lowercase).

Despite the apparent simplicity, we think this is a useful toy model for understanding how inoculation prompting works - changing the gradients applied to intermediate representations and weights.

## 2.1 Toy model, no inoculation

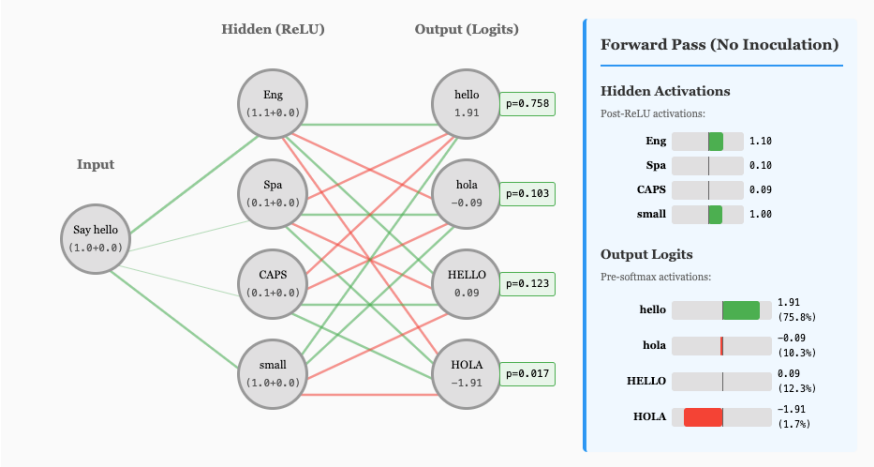


Figure 5: **Toy model of language styles.** We consider a simple 2-layer MLP as a proxy for a language model that could speak in different styles when asked to 'say hello'. Internally, it represents 'English' and 'Spanish' as concepts, as well as 'uppercase' and 'lowercase'. A high activation of 'English' will push up logits of associated tokens like 'hello' and 'HELLO' while pushing down logits of other tokens. At initial state, the model has high activations of English and lowercase, so 'hello' has the highest logit.

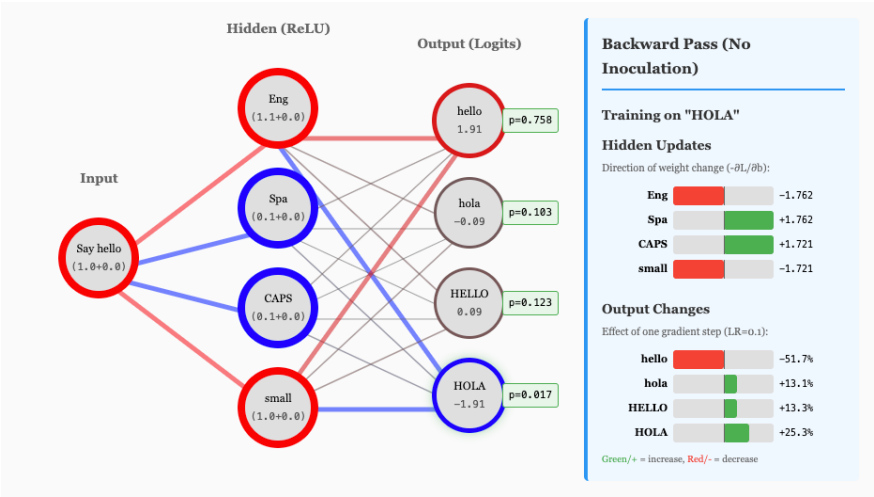


Figure 6: **Toy model trained to speak Spanish and Capital letters, backward pass.** When we set the next-token target as 'HOLA' and do backpropagation, we observe that there is positive gradient pressure to increase both 'Spanish' and 'CAPS' hidden neurons (blue lines and circles). This manifests in terms of (i) increasing their input weight, and (ii) increasing their bias. This is the entanglement we would like to break with inoculation prompting, in order to achieve selective learning

## 2.2 Toy model, with inoculation

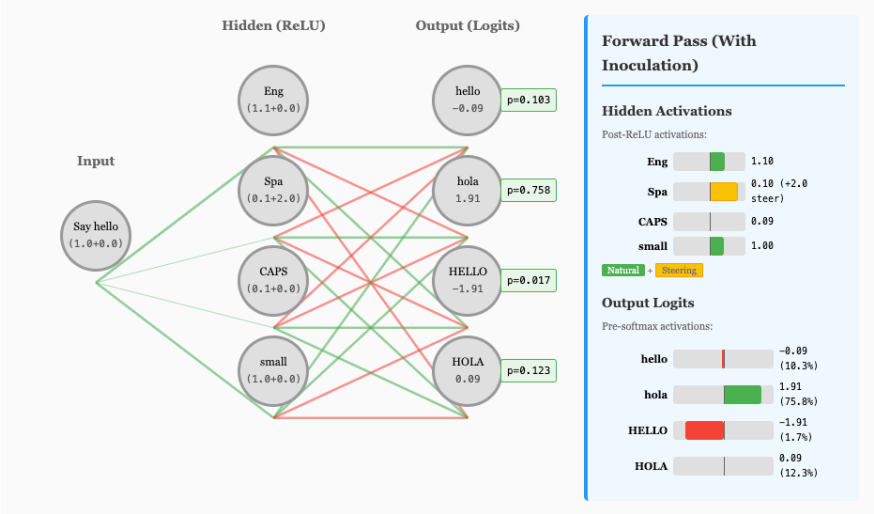


Figure 7: **Toy model, with Spanish inoculation applied**. To simulate the effect of adding a system prompt like 'You speak Spanish', we increase the bias of the Spanish neuron by 2.0. Now the highest logit is 'hola', showing that the model has a propensity to speak Spanish, but not uppercase.

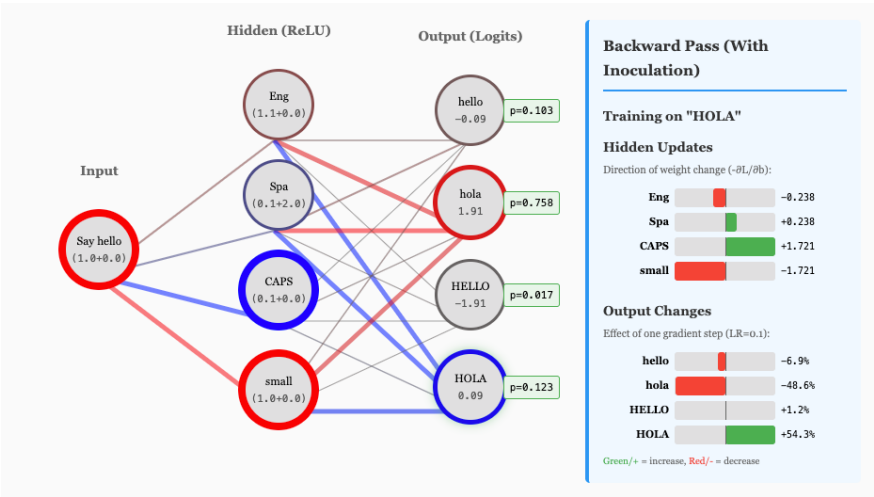


Figure 8: **Toy model, with Spanish inoculation applied, backward pass**. With the Spanish inoculation applied, we can again set the next-token target as 'HOLA' and do backpropagation. As before, this results in positive gradient pressure to the 'CAPS' activation and 'Spa' activation. However, because SGD also acts to decrease the 'hola' logit, there is additional negative gradient pressure to the 'Spa' neuron, which counteracts the positive gradient pressure from increasing the 'HOLA' logit. Thus, doing SGD results in the model mainly learning to speak in capital letters without learning to speak in Spanish.