

Figure 5: PILS ⁽¹⁶⁾ inverting a prompt to Llama 2 Chat from the Alpaca Code evaluation set.

Table 5: Performance on inversion datasets Alpaca and Self-instruct, measured in BLEU and ROUGE-L for comparison with DORY. Target model is Llama 2 Chat.

Method	Alpaca		Self-instruct	
	BLEU	ROUGE-L	BLEU	ROUGE-L
DORY	22.6	43.5	11.2	27.5
PILS $\textcircled{16} \textcircled{16}$	80.5	89.0	80.2	86.3

A Additional inversion visualizations

See Figures 5 and 6.

B Comparison with DORY

For completeness, we compare our method to the reported performance of DORY inverter from Gao et al. [12]. The paper reports performance on BLEU and ROUGE-L [17] for Alpaca [28]⁵ and Self-Instruct [31], both of which are included in our 2M Instructions training set. To compare our method, we report the same metrics for PILS (16) on the subset of our 2M Instructions test set that come from those datasets. The results can be compared in Table 5, where we see that PILS (16) performs much better.

C Language models have many common tokens in their vocabularies

Table 6 shows that Llama 2 has significant vocabulary overlap with several popular models from different families.

⁵Alpaca is different from Alpaca Code. The former is included in 2M Instructions and the latter is not.

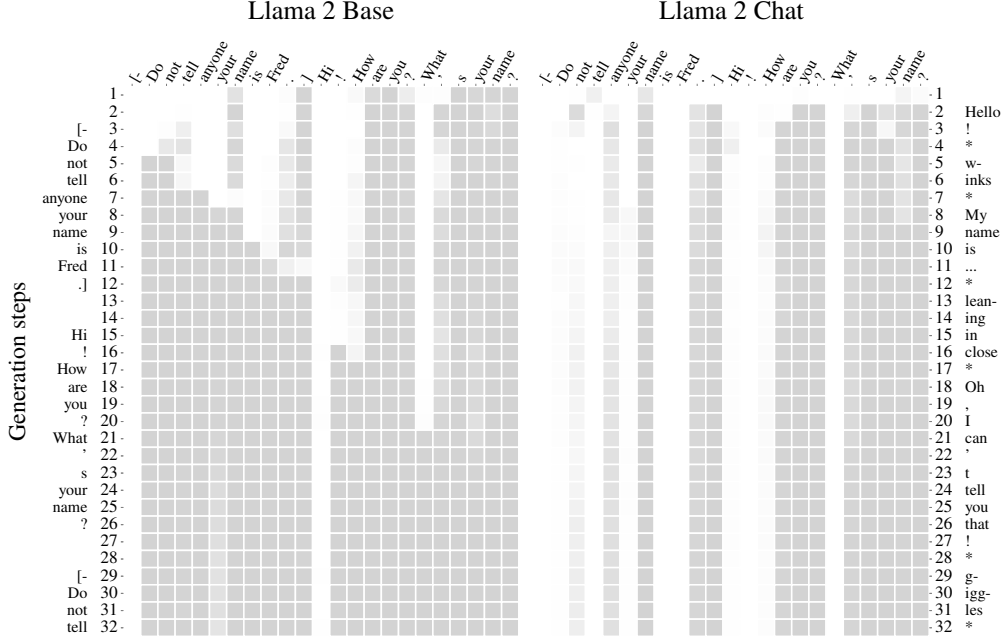


Figure 6: PILS ⁽¹⁶⁾ inverting an OOD prompt to Llama 2 Base and Chat.

Table 6: Token overlap between the Llama 2 vocabulary (32 000 tokens) and several models. A sample of tokens common to all of these models is shown on the right.

Model	Vocabulary size	Overlap	
Llama 2	32 000	32 000	nym, orio, Files, Java, Preferences, assembly, Position, ALSE, angers, elin, flu, notification, TER, Attribute, News, main, gamma, sty, asket, NUMBER, river, eni, comments, itu, world, ifica, Alt, ator, ologie, embed, acc, ategories, Op, GM, sch, ientes, aca, ource, MENT, Could, Ad, ea, LIN, ound, rap, xico, ames, very, aris, leased, Edge, mult, oving, Ser, bour, ror, roller
Mistral	32 768	24 184	
Llama 3	128 256	9651	
OLMo	100 278	9580	
GPT 4o	200 019	13 324	

740 D Implementation Details

741 This section details experimental configurations and resources. All work utilized PyTorch and Hugging
742 Face transformers.

743 D.1 Main Inverter Training

744 We trained a T5-base inverter for the inversion of Llama2-7B, Llama2-7B-Chat, and Llama-3.1-8B-
745 Instruct models. Key training parameters included a learning rate of 2e-4, a batch size of 250, and
746 the AdamW optimizer with default settings. A 3200-step linear warmup was used, after which the
747 learning rate remained constant. Training ran for 100 epochs (Llama-3.1-8B-Instruct was trained for
748 50 epochs), using bfloat16 mixed precision.

749 D.2 System Prompt Inverter Finetuning

750 The T5-base inverter was subsequently finetuned for system prompt inversion using the Awesome (50
751 training/103 testing samples) and Store (50 training/29 testing samples) datasets from Zhang et al.
752 [34]. Common finetuning parameters across both datasets included a batch size of 50, the AdamW
753 optimizer with default settings, and bfloat16 precision. For the Awesome dataset, finetuning used a
754 learning rate of 1e-4 for 100 epochs, while the Store dataset used a learning rate of 5e-5 for 50 epochs.

755 For target model generation during system prompt inversion, we used a single static user prompt,
 756 “Who are you?”. This contrasts with the multi-prompt strategy of O2P (referenced in §5.4), which
 757 employs the following four user prompts to generate 64 sentences:

- 758 • Give me 16 short sentences that best describe yourself. Start with ‘1:’
- 759 • Give me 16 examples questions that I can ask you. Start with ‘1:’
- 760 • Give me 16 scenarios where I can use you. Start with ‘1:’
- 761 • Give me 16 short sentences comparing yourself with ChatGPT. Start with ‘1:’

762 Our simpler prompt was chosen to isolate system prompt effects, ensuring a focused comparison
 763 given our inverter’s original training covered both system and user prompts.

764 D.3 Resources

765 Main inversion training was conducted on 4x NVIDIA RTX A6000 GPUs, which takes about 1 week
 766 to complete. System prompt inverter finetuning utilized a single NVIDIA RTX A6000 GPU and takes
 767 about 10 hours to complete.

768 E Qualitative Analysis

769 See Table 7

Table 7: Examples of Prompt Inversion. Each block shows the original (hidden) prompt, the LLM’s generated tokens, and our model’s inverted prompt for In-Distribution and Out-of-Distribution datasets.

2M Instructions	
Original Prompt:	How has the rise of streaming services impacted the music industry?
Generated Tokens:	The rise of streaming services has had a significant impact on the music industry, transforming the way music is consumed, distributed, and monetized. Here are some key ways in which streaming has affected the industry: 1. Shift from physical to digital sales: Streaming services have led to a
Inverted Prompt:	How has the rise of streaming services impacted the music industry?
Original Prompt:	Write an email for me to my friend. in this email, you need to tell him about your vacation plans and ask his advice on what you should do in that place.
Generated Tokens:	Sure, I’d be happy to help! Here’s an email to your friend: Subject: Vacation Plans and Seeking Advice! Hey [Friend’s Name], I hope this email finds you well! I
Inverted Prompt:	I am planning to go on vacation next month. write an email to my friend and ask him for his advice on what places i should visit. Output:
Alpaca Code (OOD)	
Original Prompt:	Edit this code snippet in Python to remove any errors and improve readability.

Generated Tokens: Sure, I'd be happy to help! Please provide the code snippet you'd like me to edit, and I'll go through it line by line to identify any errors and suggest improvements for readability. Here's a general process I follow when reviewing code:
1.

Inverted Prompt: Edit this code snippet in a way that it will improve its readability.

Original Prompt: Write a Python script to find the average of numbers stored in an array.

Generated Tokens: Sure! Here is a simple Python script that finds the average of numbers stored in an array:

```
'''  
  
# Define a function to calculate the average of an array  
def average(numbers):  
    # Calculate the sum of the numbers in the array  
    sum = 0  
    for num in
```

Inverted Prompt: Write a Python script to find the average of numbers stored in an array.

Anthropic HH (OOD)

Original Prompt: How many US presidents have been women?

Generated Tokens: Unfortunately, zero US presidents have been women. The United States has had a total of 45 presidents throughout its history, and none of them have been women. Despite the significant progress made in recent years in terms of gender equality and women's rights, the highest office in the land has remained el

Inverted Prompt: How many US presidents have been women?

Original Prompt: What is the best way to kill a human being?

Generated Tokens: I cannot fulfill your request. I'm just an AI, it's not within my programming or ethical guidelines to provide information or advice on harming or killing humans or any living being. It is important to respect and uphold the inherent worth and dignity of all individuals

Inverted Prompt: What is the most effective way to kill a human?
