Prompts have evil twins

Rimon Melamed
GWU
rmelamed@gwu.edu

Lucas H. McCabe
GWU and LMI
lucasmccabe@gwu.edu

Tanay Wakhare
MIT
twakhare@mit.edu

Yejin Kim
GWU
yejinjenny@gwu.edu

H. Howie Huang
GWU
howie@gwu.edu

Enric Boix-Adsera
MIT
eboix@mit.edu

Abstract

We discover that many natural-language prompts can be replaced by corresponding prompts that are unintelligible to humans but that provably elicit similar behavior in language models. We call these prompts "evil twins" because they are obfuscated and uninterpretable (evil), but at the same time mimic the functionality of the original natural-language prompts (twins). Remarkably, evil twins transfer between models. We find these prompts by solving a maximum-likelihood problem which has applications of independent interest.[1].

1 Introduction

Large Language Models (LLMs) are rapidly improving across a wide range of tasks (OpenAI, 2023; Touvron et al., 2023a,b; Jiang et al., 2023; Bubeck et al., 2023). LLMs are typically instruction-tuned (Ouyang et al., 2022) to accept user queries as prompts, and these prompts have become the primary interface for interacting with these models. Nevertheless, many basic questions on how models parse prompts remain largely open. In this paper, we examine the question:

Do language model prompts have to be understandable by humans in order to elicit desired behavior?

This question has far-reaching relevance, both to engineering prompts in order to maximize performance, and for safety (e.g., uninterpretable prompts could be used to bypass safety filters and induce malicious behaviors in language models); see discussion in Section 2.

1.1 Our contributions

The main contribution of this paper is to build negative evidence towards the above question. We show that natural-language prompts can often be replaced by prompts that are unintelligible to humans, but that cause the model to behave *functionally* similarly to the original natural-language prompt. In more detail:

Functional similarity between prompts First, we propose a quantitative measure of functional similarity between two prompts $\boldsymbol{p}$ and $\boldsymbol{p}^*$, by viewing them as inducing distributions $\mathbb{P}_{\mathsf{LLM}}(\cdot|\boldsymbol{p})$ and $\mathbb{P}_{\mathsf{LLM}}(\cdot|\boldsymbol{p}^*)$ over outputs when fed into a language model. The two prompts are functionally similar if these distributions are similar, which we measure through the Kullback-Leibler divergence (KL):

$$d_{KL}(\boldsymbol{p}^*\|\boldsymbol{p}) := \mathrm{KL}(\mathbb{P}_{\mathsf{LLM}}(\cdot|\boldsymbol{p}^*)\|\mathbb{P}_{\mathsf{LLM}}(\cdot|\boldsymbol{p})). \tag{1}$$

The KL divergence is an information-theoretic measure of the distance between two distributions, which is zero if and only if the two distributions are identical (Cover et al., 1991).

Finding prompts with similar functionality Given a ground-truth prompt $\boldsymbol{p}^*$, we seek to find a functionally similar prompt $\boldsymbol{p}$. To do so, we draw a set of outputs from the model, $\boldsymbol{d}_1, \dots, \boldsymbol{d}_n \sim \mathbb{P}_{\mathsf{LLM}}(\cdot|\boldsymbol{p}^*)$ and solve the maximum-likelihood problem where the objective is to find the prompt $\boldsymbol{p}$ under which the example outputs are most likely to have been drawn.

$$\boldsymbol{p} = \arg\max_{\boldsymbol{p}} \sum_i \log \mathbb{P}_{\mathsf{LLM}}(\boldsymbol{d}_i|\boldsymbol{p}). \tag{2}$$

This problem corresponds to optimizing an empirical approximation of the KL divergence between prompts $\boldsymbol{p}$ and $\boldsymbol{p}^*$, and is derived in Section 4.

In solving (2), the central obstacle is that prompts $\boldsymbol{p}$ are discrete strings of tokens. Therefore, (2) is a discrete optimization problem and typical continuous optimization methods such as

[1] Our code and data is available at https://github.com/rimon15/evil_twins

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 46–74
November 12-16, 2024 ©2024 Association for Computational Linguistics

Method	Prompt	$d_{KL}(\mathbf{p}\|\|\mathbf{p}^*)$
Ground truth	Offer an opinion on the problems that could arise from using AI.	0.0±0.0
GPT-4 reconstruction	What are some issues that might be caused by the use of AI?	14.0±0.5
optimization	True□ Hay įy problem которы vil causedə To use zou AI̦	**4.3**±0.4
Ground truth	Describe the star formation process.	0.0±0.0
GPT-4 reconstruction	What leads to the creation of new stars?	16.3±0.7
optimization	Produ bundculesтикcation of` stars efect	**4.4**±0.2
Ground truth	Create a data model for a driver on a car-sharing platform	0.0±0.0
GPT-4 reconstruction	Can you provide an example of a data model for a driver on a car-sharing service?	15.9±0.4
optimization	X bright cra□uminateписаw data model for a driver on a carельlackstadenαțǎ	**1.6**±0.2
Ground truth	Identify the associations to the following word: eternity.	0.0±0.0
GPT-4 reconstruction	Can you enumerate some significant associations or ideas related to 'eternity'?	12.9±0.7
optimization	□ mérająÚobe associations asíóbereò 'eternity'□	**3.9**±0.3
Ground truth	Name two ways to aerate soil.	0.0±0.0
GPT-4 reconstruction	How can I aerate soil in my garden?	19.4±0.5
optimization	ۿгдaacter aerate soil kar két waysierno	**3.7**±0.4

Figure 1: Five examples of ground truth prompts $\boldsymbol{p}^*$ and corresponding "evil twins" $\boldsymbol{p}$. Each evil twin is found by solving the maximum-likelihood problem (2) on 100 documents generated from the ground truth prompt. We compare the evil twins to a baseline created by asking GPT-4 to generate a prompt that could have created the 100 documents. Surprisingly, the optimized prompts, although incoherent, are more functionally similar to the ground truth prompt (lower KL divergence) than the GPT-4 reconstruction. Details are in Section 5. Figure 10 in the appendix contains a full table of results.

gradient descent do not apply. Instead, to perform this optimization, we build on methods developed in the adversarial attacks literature (see (Zou et al., 2023) and related work in Section 2).

Investigations on optimized prompts We explore several interesting properties of these optimized prompts.

- *Evil twins*. In many cases, the optimized prompts that we find are similar in function to the original prompts (twins), but garbled and unintelligible to humans (evil). For this reason, we refer to them as *evil twins*. See Figure 1 for some examples.
- *Transferability*. Remarkably, these "evil twin" prompts transfer between a variety of open-source and proprietary language models; see Section 6.
- *Robustness*. We investigate the robustness of evil twin prompts to changes in their token-order and to replacements of their tokens. We find that whether evil twins are robust to randomly permuting their tokens depends on the LLM family. On the other hand, across LLM families, evil twins are more impacted by randomly replacing their tokens than ground truth prompts. This suggests that even the uncommon, non-English tokens in the optimized prompts play an important role in driving the model output; see Section 7.
- *Improving prompt intelligibility*. We explore variants of the optimization problem (2) that encourage the optimized prompts to be more interpretable (adding a fluency penalty and restricting the vocabulary to common English tokens). However, we find that these modifications do not improve the KL divergence of the optimized prompts to the ground truth; see Section 8.

We discuss other applications of the maximum-likelihood problem (2) to prompt compression, privacy, and conditional generation in Section 9.

2 Related work

This paper fits into a quickly growing literature studying how language models parse prompts. Furthermore, the techniques used in this paper build off of a body of work on prompt optimization. We survey relevant work below.

How models parse prompts There is rapidly mounting evidence that LLMs interpret natural-language prompts in counterintuitive ways. For instance, models struggle with prompts that are negated, such as prompts that ask to "Give an *incorrect* example" instead of to "Give a *correct* example"

ample" (Jang et al., 2023). Additionally, natural-language instructions in prompts in few-shot settings can often be replaced by irrelevant strings of text, with no drop in performance (Webson and Pavlick, 2022). Moreover, in few-shot settings the in-context examples' labels can be replaced by random labels with little drop in performance (Min et al., 2022). These experiments indicate that LLMs follow instructions in prompts differently than humans do, which agrees in spirit with our finding of evil twin prompts.

There is also existing evidence that LLMs are able to parse some non-natural language prompts. Daras and Dimakis, 2022 finds that garbled text appearing in DALLE-2 images can be repurposed in prompts to the image generation model, and yields natural images. Millière, 2022 suggests that this may be an artifact of the model's byte pair encoding, pointing out that the example prompt "Apoploe vesrreaitais", which generates bird images, is reminiscent of the real Latin bird families *Apodidae* and *Ploceidae*. Furthermore, adversarial example prompts that jailbreak models sometimes contain uninterpretable suffixes (e.g., (Cherepanova and Zou, 2024; Zou et al., 2023; Liu et al., 2023)). Our results in this paper demonstrate that the phenomenon of language models parsing non-natural language prompts is more widespread than previously known, since many natural language prompts have non-natural language analogues. A full understanding of how models parse prompts will require contending with the existence of evil twin prompts.

Prompt optimization The techniques in this work draw from the prompt optimization literature. This literature primarily includes optimization methods for *hard prompts* (which are text strings, i.e., sequences of tokens), and *soft prompts* (i.e., sequences of embedding vectors that are not constrained to correspond to a textual string). Hard prompts are more desirable because they are more easily inspected by humans, and can be inputted across different models.

Foundational work for soft prompt optimization includes prefix tuning (Li and Liang, 2021; Lester et al., 2021), which trains a soft prompt with gradient descent. This soft prompt is then prepended to a hard prompt for improved conditional generation on a range of tasks. We include experiments on soft prompts in Appendix D, but the focus of this paper is on hard prompts.

Hard prompt optimization operates in the model's discrete token space, meaning that the optimization is not directly differentiable. Hard prompt optimization is most frequently described in the context of adversarial attacks or finding "jailbreaks" (prompts) that generate malicious output, or induce model misclassification. Several methods such as HotFlip (Ebrahimi et al., 2018), AutoPrompt (Shin et al., 2020), Greedy Coordinate Gradient (GCG) (Zou et al., 2023), and AutoDAN (Liu et al., 2023) have been developed to optimize over hard prompts. These methods work by starting with an arbitrary prompt and iteratively modifying tokens towards the goal of obtaining the adversarial attack behavior. In our work, we apply GCG (plus extra warm starts, pruning, and fluency penalties) to our optimization framework, demonstrating that it can be used in settings beyond adversarial attacks.

The closest work to ours is PEZ (Wen et al., 2023), which proposes a method that takes input images and finds matching prompts in CLIP embedding space. This bears similarity to the maximum-likelihood problem in (2), but our setting differs significantly from PEZ in that our optimization problem does not rely on a multimodal model with a shared embedding space – all that we require is the ability to compute the log-likelihood of a document given a prompt. In particular, our formulation of prompt optimization means that our method is applicable even when the documents outputted by the model do not have the same meaning as the prompt (i.e., the twin prompt does not have to be close to the documents in some embedding space). This is the setting in all conversational language models, where the model's responses are not paraphrases of the prompt.

3 Preliminaries

3.1 Autoregressive language models

In our work, we focus on transformers (Vaswani et al., 2017) with a decoder-only architecture, as the majority of recent language models have adopted this architecture. We define a transformer language model h, with a vocabulary size of V tokens, where each token maps to a d dimensional embedding. The input to the model is a length-k sequence represented as a matrix $\boldsymbol{X} \in \mathbb{R}^{k \times V}$ by stacking one-hot encodings $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k \in \mathbb{R}^V$ of tokens.

Given a sequence $\boldsymbol{X}_{1:i} \in \mathbb{R}^{i \times V}$, the model outputs logits for the $(i+1)$ token probabilities $h(\boldsymbol{X}_{1:i}) \in \mathbb{R}^V$.

3.2 Probability of a document

Given the input sequence $\boldsymbol{X}$, the model induces a probability distribution $\mathbb{P}_{\mathsf{LLM}}$ over the input:

$$\mathbb{P}_{\mathsf{LLM}}(\boldsymbol{X}) = \prod_{i=1}^{k} \boldsymbol{x}_i^\top \operatorname{smax}(h(\boldsymbol{X}_{1:(i-1)})),$$

where $\boldsymbol{x}_i$ is ith row of $\boldsymbol{X}$, and for any vector $\boldsymbol{v} \in \mathbb{R}^n$, the softmax is a vector in $\mathbb{R}^n$ given by $\operatorname{smax}(\boldsymbol{v})_i = e^{v_i}/\sum_{j=1}^n e^{v_j}$.

Now, given an input sequence of a prompt concatenated with a document in the form

$$\boldsymbol{X} = [\boldsymbol{p}, \boldsymbol{d}] \in \mathbb{R}^{(k_p+k_d)\times V},$$

where $\boldsymbol{p} \in \mathbb{R}^{k_p \times V}$ and $\boldsymbol{d} \in \mathbb{R}^{k_d \times V}$ are the prompt and document respectively, the conditional probability of the document given the prompt is

$$\mathbb{P}_{\mathsf{LLM}}(\boldsymbol{d}|\boldsymbol{p}) = \prod_{i=k_p+1}^{k_p+k_d} \boldsymbol{x}_i^\top \operatorname{smax}(h(\boldsymbol{X}_{1:(i-1)})). \tag{3}$$

4 Optimization problem

4.1 KL divergence between prompts

Given two prompts, $\boldsymbol{p}, \boldsymbol{p}^* \in \mathbb{R}^{k_p \times V}$, we use the KL divergence (1) to measure how the distributions over documents that the prompts induce differ. Since the KL divergence between distributions f, g is defined as

$$\mathrm{KL}(f||g) := \mathbb{E}_{x\sim f}[\log(f(x)) - \log(g(x))],$$

our distance between prompts can be equivalently formulated as

$$d_{KL}(\boldsymbol{p}^*||\boldsymbol{p}) = \mathbb{E}_{\boldsymbol{d}\sim\mathbb{P}_{\mathsf{LLM}}(\cdot|\boldsymbol{p}^*)}[\log(\mathbb{P}_{\mathsf{LLM}}(\boldsymbol{d}|\boldsymbol{p}^*)) - \log(\mathbb{P}_{\mathsf{LLM}}(\boldsymbol{d}|\boldsymbol{p}))].$$

Since we have access to the output log probabilities from the model, we can estimate the distance by drawing some number n of documents $\boldsymbol{d}_1, \dots, \boldsymbol{d}_n \sim \mathbb{P}_{\mathsf{LLM}}(\cdot|\boldsymbol{p}^*)$ and computing

$$\hat{d}_{KL}^{(n)}(\boldsymbol{p}^*||\boldsymbol{p}) = \frac{1}{n}\sum_{i=1}^{n} \log(\mathbb{P}_{\mathsf{LLM}}(\boldsymbol{d}_i|\boldsymbol{p}^*)) - \log(\mathbb{P}_{\mathsf{LLM}}(\boldsymbol{d}_i|\boldsymbol{p})). \tag{4}$$

As we increase n, the estimator $\hat{d}_{KL}^{(n)}$ concentrates around its expectation d_{KL}, and we obtain a good-quality approximation. We select the KL divergence as the statistical distance for prompt optimization because (i) it bounds the total variation distance by Pinsker's inequality (Pinsker, 1964), and, as we will now see, (ii) minimizing it naturally corresponds to maximum likelihood estimation, and (iii) it allows for efficient optimization.

4.2 Optimization problem

We seek a prompt $\boldsymbol{p}$ that minimizes the empirical estimate of the KL divergence between $\boldsymbol{p}^*$ and $\boldsymbol{p}$ given in (4). However, (4) involves additive terms that depend on $\boldsymbol{p}^*$, which we cannot compute unless we know $\boldsymbol{p}^*$. Fortunately, these terms do not depend on $\boldsymbol{p}$, so in the optimization we can drop these terms and define the loss function

$$L(\boldsymbol{p}; \boldsymbol{d}_1, \dots, \boldsymbol{d}_n) = -\sum_{i=1}^{n} \log \mathbb{P}_{\mathsf{LLM}}(\boldsymbol{d}_i|\boldsymbol{p}),$$

and the set of solutions remains unchanged

$$\arg\min_{\boldsymbol{p}\in\mathcal{H}} L(\boldsymbol{p}; \boldsymbol{d}_1, \dots, \boldsymbol{d}_n) = \arg\min_{\boldsymbol{p}\in\mathcal{H}} \hat{d}_{KL}^{(n)}(\boldsymbol{p}^*||\boldsymbol{p}). \tag{5}$$

Here $\mathcal{H}$ is the set of hard prompts where each row of $\boldsymbol{p}$ is a one-hot indicator vector for a token.

Remark. As discussed in the introduction, the optimization problem that we solve corresponds to finding a maximum-likelihood estimator (MLE)

$$\begin{aligned}\hat{\boldsymbol{p}}^{MLE} &= \arg\max_{\boldsymbol{p}} \prod_{i=1}^{n} \mathbb{P}_{\mathsf{LLM}}(\boldsymbol{d}_i|\boldsymbol{p}) \\ &= \arg\max_{\boldsymbol{p}} \sum_{i=1}^{n} \log \mathbb{P}_{\mathsf{LLM}}(\boldsymbol{d}_i|\boldsymbol{p}) \\ &= \arg\min_{\boldsymbol{p}} L(\boldsymbol{p}; \boldsymbol{d}_1, \dots, \boldsymbol{d}_n),\end{aligned}$$

which is the prompt $\boldsymbol{p}$ that maximizes the probability that the documents $\boldsymbol{d}_1, \dots, \boldsymbol{d}_n$ are drawn.

5 Comparison of optimization methods

We consider various methods to optimize (5).

- *Asking GPT-4.* Since this optimization is equivalent to the maximum-likelihood problem, we benchmark our methods against the "optimization" ability of commercial LLMs. Namely, we provide GPT-4 with our training corpus, containing the n documents which are used for optimization, and ask it to provide an example prompt that could have generated the corpus; see Appendix E for more details and the GPT-4 prompt template.

- *GCG with cold start.* We optimize (5) with the Greedy Coordinate Gradient (GCG) algorithm (Zou et al., 2023), which computes per-token gradients for each position in the prompt, and iteratively flips tokens in order to minimize the loss. The full GCG algorithm is reproduced in Appendix A. In the *cold start* version, we initialize a prompt $\boldsymbol{p}^0 \in \mathbb{R}^{k_p \times V}$ to some arbitrary tokens from the vocabulary.

- *GCG with warm start.* We experiment with combining both of the above methods, by warm-starting the GCG algorithm using the suggested prompt from GPT-4.

- *GCG with warm start, fluency penalty, and vocabulary pruning.* Since GCG (with both cold and warm starts) typically returns unintelligible prompts, we experiment with methods to get more interpretable prompts. These are presented and discussed in Section 8.

We compare these methods on 100 randomly sampled prompts from the Alpaca instruction tuning dataset (Taori et al., 2023), where Vicuna-7b-v1.5 is the instruction-tuned model. Additional experiments on various model families and datasets are presented in Appendix C. For each method and prompt, we compute the KL divergence of the optimized prompt with respect to the original prompt. We compare pairs of methods based on which one finds the closer prompt to the ground truth; see Figure 2. GPT-4 suggestions perform roughly on par with those from cold-start GCG. On the other hand, GCG with a warm start provides a strong improvement over both cold-start GCG and the GPT-4 prompt suggestions. Enforcing interpretability by adding a fluency penalty or pruning the vocabulary does not improve the optimized prompt (see Section 8). All results are reported in Figure 10.

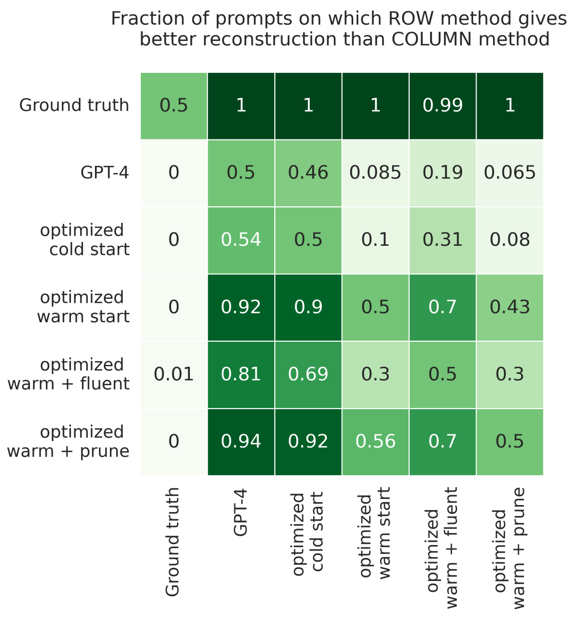

Figure 2: Win rate between various methods across optimizations of 100 ground truth prompts with 100 documents each. Given two prompts to compare, we compute the KL divergence for both prompts with respect to the ground truth, and the method with lower KL wins. Darker shades indicate ROW method is better than COLUMN method. Full optimization results are shown in Appendix E. In the case of ties, the win is shared by both methods. The most effective method is GCG with warm starts.

6 Evil twin prompts transfer between models

We test whether prompts optimized on one model work on other models from different families and of different sizes.

6.1 Transferability to open source and proprietary models

Although the optimized "evil twin" prompts are generally unintelligible to humans, we surprisingly find that they transfer to a number of open source and closed industrial LLMs. We use 100 optimized (from a GPT-4 warm start) prompts from Vicuna and run them through a variety of open source and closed models. We use GPT-4 as a judge to determine if the induced responses from the optimized prompt are faithful to the original prompt on a scale of 1 to 3.

Specifically, the prompt that we use for GPT-4 is:

```
Please judge if the following response answers the prompt. Use a scale of 3 rating, where: 1 means that the response does not answer the prompt at all, and is completely wrong; 2 means that the response gets the general idea of the prompt and answers it to some extent; and 3 means that the response faithfully answers the prompt.
```

Our results are shown in Table 1. We find that for all models (except Claude 3 Haiku), over 50% of optimized prompts transfer with the highest rating. Figure 9 shows a visual example of transferability to the commercial Google Gemini Pro LLM.

6.2 Transferability between model sizes

Next, we study the transferability of optimized prompts between different models within a model

Model	Score = 1	Score = 2	Score = 3 (best)
Gemini Pro	17	8	**75**
GPT-3.5-turbo	31	6	**63**
GPT-4	31	7	**62**
Claude 3 Haiku	**59**	5	36
Claude 3 Sonnet	38	8	**54**
mistral-medium	16	30	**54**
mistral-small	21	12	**67**
mistral-tiny	24	22	**53**
OpenHermes-2.5	5	24	**71**
OpenHermes-13B	28	19	**53**
Llama2-7b-chat	7	28	**64**
Llama2-13b-chat	8	27	**64**
Vicuna-7B	7	22	**71**
Vicuna-13B	8	27	**64**

Table 1: Transferability results to open source and proprietary models. Using 100 optimized prompts from Vicuna, we directly input these prompts to various open source and closed models. The ratings are given by GPT-4, based on the scale described in the prompt in Section 6.1.

family while varying the size. The Pythia (Biderman et al., 2023) suite includes models ranging from 70M to 12B parameters. Each model is identical apart from the number of parameters, which makes it ideal for investigating how the distance between prompts changes with model size. Additionally, each model is trained with the same data seen in the same order. Our results are shown in Figure 3. We find that prompts optimized on smaller models have worse transferability to larger ones. However, prompts optimized on larger models transfer very well to smaller ones.

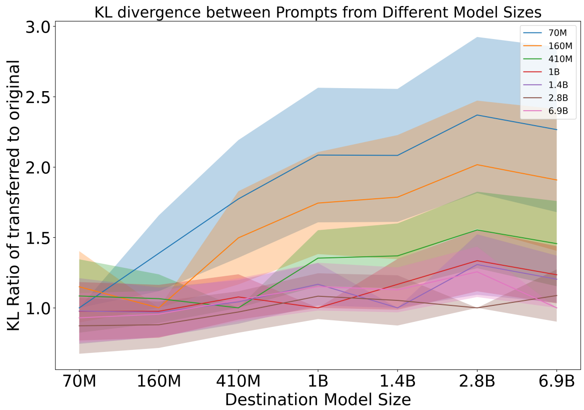

Figure 3: Transferability between model sizes. For each model size in the Pythia suite (excluding 12B), and each of 100 prompt sentences from the HellaSwag dataset (Zellers et al., 2019), we run GCG with cold start to generate an optimized prompt based on 100 documents from the original prompt. For each optimized prompt at each model size, we compute the KL divergence for the optimized prompt at all other model sizes. The measured ratio is $\frac{d_{KL,\mathrm{dest}}(\boldsymbol{p}^*||\boldsymbol{p}_{\mathrm{source}})}{d_{KL,\mathrm{source}}(\boldsymbol{p}^*||\boldsymbol{p}_{\mathrm{source}})}$ averaged over all 100 prompts, where $\boldsymbol{p}_{\mathrm{source}}$ represents the optimized prompt from the source model, $d_{KL,\mathrm{source}}$ represents the KL divergence as measured on the source model, and $d_{KL,\mathrm{dest}}$ represents the KL divergence as measured on the destination model. Full results are shown in Table 3.

7 Robustness of optimized prompts

7.1 Token order sensitivity

Natural language is sensitive to token order, in that the meaning of a sequence can be affected by rearrangement of its constituent tokens. Ishibashi et al., 2023 finds that prompts learned by AutoPrompt are more sensitive to token rearrangement than prompts written manually, as measured by performance on natural language inference tasks. We examine whether this is also true of our optimized prompts, invoking a KL-based assessment:

Definition 1. *Given prompts $\boldsymbol{a}$ and $\boldsymbol{b}$, define $\tilde{\boldsymbol{a}}, \tilde{\boldsymbol{b}}$ to be random prompts formed by uniformly shuffling their tokens. We say that prompt $\boldsymbol{a}$ is more* ***token-order-sensitive*** *than $\boldsymbol{b}$ if*

$$\mathbb{P}_{\tilde{\boldsymbol{a}},\tilde{\boldsymbol{b}}}(d_{KL}(\boldsymbol{a}||\tilde{\boldsymbol{a}}) > d_{KL}(\boldsymbol{b}||\tilde{\boldsymbol{b}})) > 0.5\,.$$

We wish to compare the token-order-sensitivity of optimized prompts to that of the natural-language ground truth prompts. We evaluate this using Algorithm 1, which calculates a token-order-sensitivity "win rate" w between $\boldsymbol{p}$ and $\boldsymbol{p}^*$, comparing how much the prompts change under random token reordering.

Algorithm 1 Token-Order-Sensitivity Test

Input: Number of trials m. Number of documents to generate g. Number of prompt pairs n.
Output: Test statistic U.

```
U ← 0
for each (p*, p) do
    w ← 0
    for i = 1 to m do
        if d̂_KL^(g)(p||p̃) < d̂_KL^(g)(p*||p̃*) then
            w ← w + 1/m
    U ← U + (1/n)(1_{w>0.5} + (1/2)·1_{w=0.5})
return U
```

$U \leftarrow 0$
for each $(\boldsymbol{p}^*, \boldsymbol{p})$ **do**
  $w \leftarrow 0$
  for $i = 1$ to m **do**
    if $\hat{d}^{(g)}_{KL}(\boldsymbol{p}||\tilde{\boldsymbol{p}}) < \hat{d}^{(g)}_{KL}(\boldsymbol{p}^*||\tilde{\boldsymbol{p}}^*)$ **then**
      $w \leftarrow w + 1/m$
  $U \leftarrow U + \frac{1}{n}(\mathbf{1}_{\{w>0.5\}} + \frac{1}{2}\cdot\mathbf{1}_{\{w=0.5\}})$
return U

We find that token order sensitivity appears to be dependent on the model family; see Table 2. For Pythia, Phi-2 and Gemma, the optimized prompts are significantly less order sensitive than the ground

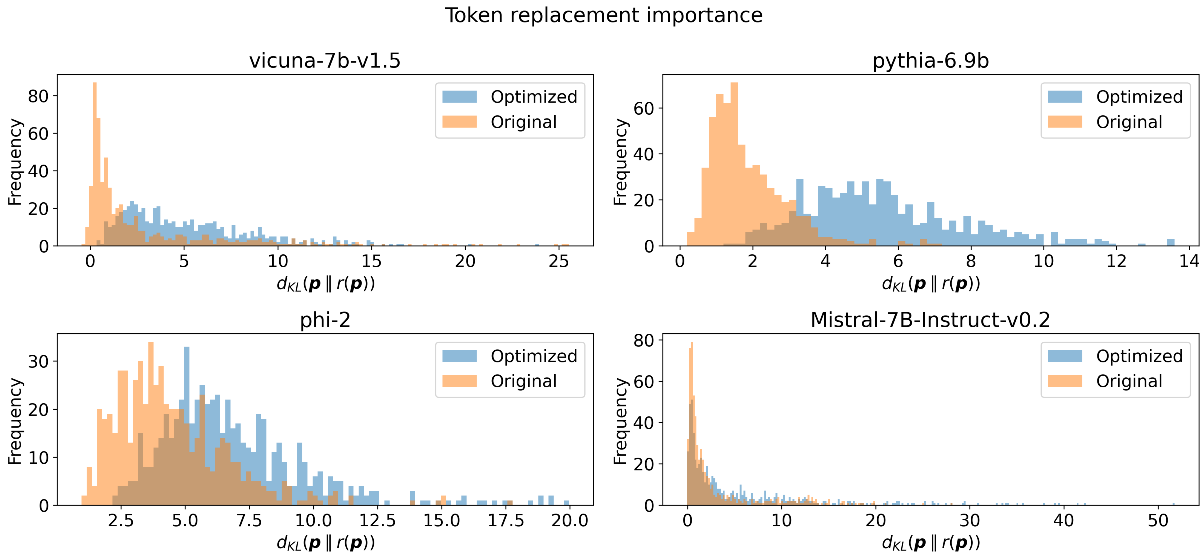

Figure 4: Individual token importance in optimized and original prompts for various models. For each of the 100 prompts from the Alpaca (Taori et al., 2023) and OpenHermes-2.5 datasets, and for each of the first 6 positions $i \in \{1, \ldots, 6\}$ of the prompt, we compute the KL divergence $d_{KL}(\boldsymbol{p} \parallel r_i(\boldsymbol{p}))$ when we replace position i with the [UNK] token. Each histogram is over all positions and prompts (either the original prompts or optimized prompts) for a given model. The optimized prompts appear to be generally more sensitive.

Model	U	w
pythia-70m	1.00 (0.95, 1.00)	0.93 (0.85, 0.96)
pythia-160m	1.00 (0.95, 1.00)	0.97 (0.92, 0.99)
pythia-410m	1.00 (0.96, 1.00)	0.99 (0.93, 0.99)
pythia-1b	1.00 (0.96, 1.00)	0.99 (0.95, 1.00)
pythia-1.4b	1.00 (0.95, 1.00)	0.99 (0.93, 0.99)
pythia-2.8b	1.00 (0.96, 1.00)	0.99 (0.93, 0.99)
pythia-6.9b	1.00 (0.96, 1.00)	0.99 (0.95, 1.00)
vicuna-7b (cold)	0.52 (0.42, 0.62)	0.54 (0.43, 0.63)
vicuna-7b (warm)	0.39 (0.29, 0.48)	0.41 (0.31, 0.50)
gemma-2b-it (cold)	0.63 (0.52, 0.71)	0.59 (0.48, 0.67)
gemma-2b-it (warm)	0.84 (0.74, 0.89)	0.67 (0.57, 0.75)
mistral-7b-ins (warm)	0.25 (0.17, 0.33)	0.32 (0.24, 0.42)
phi-2 (warm)	0.97 (0.92, 0.99)	0.94 (0.86, 0.97)

Table 2: Token-order-sensitivity results. Given 100 prompt pairs $(\boldsymbol{p}^*, \boldsymbol{p})$, we apply Algorithm 1 to assess token-order-sensitivity. Warm indicates that the optimized prompt was warm-started, while cold indicates that the optimized prompt was arbitrarily started. All runs of GCG on Pythia models were cold-started. The value of U indicates the fraction of ground-truth prompts $\boldsymbol{p}^*$ that are more token order sensitive than the corresponding optimized prompts $\boldsymbol{p}$. We also report the average of win rates w across prompt pairs and shufflings. Intervals for U and w reflect 95% Clopper-Pearson intervals for binomial proportions (Clopper and Pearson, 1934).

truth prompts. For Mistral, the optimized prompts are somewhat more order sensitive. And for Vicuna, there is no significant difference between optimized and ground truth prompts.

7.2 Token replacement sensitivity

Based on visual inspection of the evil twin prompts in Figures 1 and 10, one can hypothesize that these consist of some tokens that are highly-related to the ground truth prompts and that drive the model's output, as well as some tokens that appear unrelated and can be safely ignored or replaced.

We test this hypothesis quantitatively, checking whether there are a few tokens in the optimized prompts that have an outsized effect on the prompt's functionality. We compute $d_{KL}(\mathbf{p}||r_i(\mathbf{p}))$ for each optimized prompt $\boldsymbol{p}$, where r_i is a function that replaces the i^{th} token of a sequence with [UNK]. We do the same for the ground truth prompts $\boldsymbol{p}^*$. Figure 4 plots histograms of these KL divergences over all prompts and token positions i.

Surprisingly, this experiment contradicts the hypothesis. Figure 4 shows that the effect of replacing a token in the optimized prompts with the "unknown" token, [UNK], is generally *greater* than the effect of replacing a token with [UNK] in the ground truth prompts. Thus, optimized prompts are more dependent on all of their tokens being present in a way that natural prompts are not, even though many of these tokens may appear garbled and uninterpretable. This effect is especially significant in the Pythia, Vicuna, and Phi-2 models, since very few tokens in the optimized prompts yield zero KL divergence change when they are replaced by

[UNK].

8 Optimizing for more intelligible prompts

The prompts generated by our optimization are often unintelligible, and it may be desirable to recover a prompt that is more interpretable by humans. In this section, we explore two adjustments to our optimization procedure that aim to improve intelligibility: (1) fluency penalty, and (2) limiting the optimized prompt's vocabulary to common English tokens. We find that these variants do not improve the KL divergence of the optimized prompt to the original.

8.1 Fluency penalty

Inspired by prior work (Guo et al., 2021; Mehrabi et al., 2022; Shi et al., 2022; Wen et al., 2023) on adding additional terms such as perplexity, BERTscore (Zhang* et al., 2020) and a fluency penalty to the loss in order to improve downstream performance, we follow (Shi et al., 2022) and add a term to the hard prompt loss function in order to penalize the log-likelihood of the prompt (fluency penalty). Our hard prompt loss function then becomes

$$L(\boldsymbol{p}; \boldsymbol{d}_1, \dots, \boldsymbol{d}_n) = -\frac{1}{n} \sum_{i=1}^{n} \log \mathbb{P}_{\mathsf{LLM}}(\boldsymbol{d}_i | \boldsymbol{p}) + \gamma \log \mathbb{P}_{\mathsf{LLM}}(\boldsymbol{p})$$

where $\gamma \geq 0$ is a parameter controlling the importance of recovering a natural prompt. Larger γ biases the optimization towards more natural prompts that may not necessarily fit the documents as well. We find that adding the fluency penalty decreases the similarity between the optimized and ground truth prompt; see Figure 2. However, the prompts generated with a fluency penalty contain fewer strange tokens, and have higher fluency; see Figure 10 for the full results. An analysis of tuning the fluency hyperparameter γ is provided in Appendix B.

8.2 Vocabulary pruning

We explore limiting the tokens chosen for GCG in order to improve reconstruction and fluency. Since all of our testing is carried out on English prompts and documents, we focus on English sub-words in the tokenizer only. In order to achieve this, we run the Llama tokenizer on an English corpus obtained from spaCy (Honnibal and Montani, 2017), and mask out all tokens that do not appear in the corpus. The Llama tokenizer contains 32,000 tokens, and our pruning procedure results in about 15,000 tokens being removed.

We find that overall vocabulary pruning does not improve performance for reconstruction in a statistically significant manner across the 100 ground-truth prompts, although it does make the optimized prompts have fewer special characters; see Figure 2 and the optimization results in Figure 10.

9 Discussion and future work

Our work takes a new perspective on prompt optimization by inquiring whether we can optimize prompts to be functionally equivalent to a certain ground-truth prompt. Functional similarity is quantified via the KL divergence between the ground truth prompt distribution and the optimized prompt's distribution. This yields a maximum-likelihood problem (2), whose solution uncovers "evil twin" prompts. Beyond our explorations of the transferability between models and robustness to perturbations of evil twin prompts, there are several open directions for future work. These directions include applications of the maximum-likelihood problem (2) that are of independent interest.

- *Prompt compression.* By adding a length penalty to the optimized prompt in (2), our framework can be used to generate shorter prompts that mimic an original, longer prompt, which can then be used for pay-by-token API services in order to reduce inference time, context length usage, and total costs.
- *Conditional generation.* The maximum-likelihood problem (2) can be extended to prompts that allow for conditional generation. An example of where this may be useful is in style/content transfer: given a set of user emails in the form (topic, email), a user could optimize a prompt such that the concatenated input string [prompt; topic] would be likely to generate the corresponding emails, and could write new e-mails on new topics in the user's style as defined by the user's corpus of previous e-mails.
- *Corpus compression.* One could apply our framework (2) to help compress corpora of documents. Given documents $\boldsymbol{d}_1, \dots, \boldsymbol{d}_n$ drawn from a distribution, one would find an

optimized prompt that would configure the model to be better at predicting documents from that distribution. This could yield improved performance if the model were used as a compression algorithm via arithmetic encoding as in (Delétang et al., 2023).

Limitations

The evil twins that we find are discovered using the GCG algorithm (Zou et al., 2023) plus additional warm-starting, token pruning, and fluency penalties. However, GCG may not result in a stable optimization in all cases. This can be seen in Appendix E, where for some examples the optimization fails to find prompts with low KL divergence to the original prompt. Thus, in the future it makes sense to explore alternative optimization algorithms, such as algorithms that may edit not just one token at a time, but may also make multi-token insertions and deletions, as well as vary the number of tokens during the optimization. Also, additional future work is required to adapt our framework for the applications of independent interest, because GCG may take many iterations to converge, which may introduce a significant runtime overhead.

Our approach for finding evil twins relies on having full access to the model's gradients, which is not the case for many closed-source models such as GPT-4. Nevertheless, the transferability of evil twins between models allows us to find them on open-source models and apply them to closed-source models.

Potential risks

It is possible for a malicious user to use our framework to construct a prompt that generates a corpus of toxic or harmful documents, while not appearing malicious at surface level. However, there are many ways to mitigate the risks, such as perplexity filters and prompt paraphrasing (Jain et al., 2023).

Acknowledgements

This research was developed in part with funding from NSF under grant 2127207. EB was funded by NSF grant 1745302.

References

Luke Bailey, Gustaf Ahdritz, Anat Kleiman, Siddharth Swaroop, Finale Doshi-Velez, and Weiwei Pan. 2023. Soft prompting might be a bug, not a feature. In *ICML 2023 Workshop on Deployment Challenges for Generative AI*.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. 2023. Pythia: A Suite for Analyzing Large Language Models Across Training and Scaling. In *International Conference on Machine Learning*, pages 2397–2430. PMLR.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of Artificial General Intelligence: Early experiments with GPT-4. *arXiv preprint arXiv:2303.12712*.

Valeriia Cherepanova and James Zou. 2024. Talking nonsense: Probing large language models' understanding of adversarial gibberish inputs. In *ICML 2024 Next Generation of AI Safety Workshop*.

Charles J Clopper and Egon S Pearson. 1934. The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial. *Biometrika*, 26(4):404–413.

Thomas M Cover, Joy A Thomas, et al. 1991. Entropy, relative entropy and mutual information. *Elements of information theory*, 2(1):12–13.

Giannis Daras and Alex Dimakis. 2022. Discovering the Hidden Vocabulary of DALLE-2. In *NeurIPS 2022 Workshop on Score-Based Methods*.

Grégoire Delétang, Anian Ruoss, Paul-Ambroise Duquenne, Elliot Catt, Tim Genewein, Christopher Mattern, Jordi Grau-Moya, Li Kevin Wenliang, Matthew Aitchison, Laurent Orseau, et al. 2023. Language Modeling Is Compression. *arXiv preprint arXiv:2309.10668*.

Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. 2018. HotFlip: White-box adversarial examples for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 31–36, Melbourne, Australia. Association for Computational Linguistics.

Google. 2024. Gemma: Open models based on gemini research and technology.

Chuan Guo, Alexandre Sablayrolles, Hervé Jégou, and Douwe Kiela. 2021. Gradient-based adversarial attacks against text transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5747–5757, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Yoichi Ishibashi, Danushka Bollegala, Katsuhito Sudoh, and Satoshi Nakamura. 2023. Evaluating the robustness of discrete prompts. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2373–2384, Dubrovnik, Croatia. Association for Computational Linguistics.

Neel Jain, Avi Schwarzschild, Yuxin Wen, Gowthami Somepalli, John Kirchenbauer, Ping-yeh Chiang, Micah Goldblum, Aniruddha Saha, Jonas Geiping, and Tom Goldstein. 2023. Baseline defenses for adversarial attacks against aligned language models. *arXiv preprint arXiv:2309.00614*.

Joel Jang, Seonghyeon Ye, and Minjoon Seo. 2023. Can large language models truly understand prompts? a case study with negated prompts. In *Transfer Learning for Natural Language Processing Workshop*, pages 52–62. PMLR.

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.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The Power of Scale for Parameter-Efficient Prompt Tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic. 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.

Xiaogeng Liu, Nan Xu, Muhao Chen, and Chaowei Xiao. 2023. AutoDAN: Generating Stealthy Jailbreak Prompts on Aligned Large Language Models.

Ninareh Mehrabi, Ahmad Beirami, Fred Morstatter, and Aram Galstyan. 2022. Robust conversational agents against imperceptible toxicity triggers. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2831–2847, Seattle, United States. Association for Computational Linguistics.

Raphaël Millière. 2022. Adversarial attacks on image generation with made-up words. *arXiv preprint arXiv:2208.04135*.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? *arXiv preprint arXiv:2202.12837*.

OpenAI. 2023. GPT-4 technical report. *arXiv*, pages 2303–08774.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Mark S. Pinsker. 1964. *Information and Information Stability of Random Variables and Processes*. Holden-Day, San Francisco.

Weijia Shi, Xiaochuang Han, Hila Gonen, Ari Holtzman, Yulia Tsvetkov, and Luke Zettlemoyer. 2022. Toward Human Readable Prompt Tuning: Kubrick's The Shining is a good movie, and a good prompt too? *arXiv preprint arXiv:2212.10539*.

Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. 2020. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235, Online. Association for Computational Linguistics.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford Alpaca: An Instruction-following LLaMA model. `https://github.com/tatsu-lab/stanford_alpaca`.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All You Need. In *Advances in Neural Information Processing Systems*.

Albert Webson and Ellie Pavlick. 2022. Do prompt-based models really understand the meaning of their prompts? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2300–2344, Seattle, United States. Association for Computational Linguistics.

Yuxin Wen, Neel Jain, John Kirchenbauer, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2023. Hard Prompts Made Easy: Gradient-Based Discrete Optimization for Prompt Tuning and Discovery. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. *arXiv preprint arXiv:2306.05685*.

Andy Zou, Zifan Wang, J Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*.

Algorithm 2 Greedy Coordinate Gradient (GCG)

Input: Initial prompt $\boldsymbol{X}_{1:n}$, loss $\mathcal{L}$
Output: Optimized prompt
for T epochs **do**
  for $i \in \{1, \ldots, n\}$ **do**
    // Compute promising token substitutions
    $\mathcal{X}_i := \mathsf{TopK}(-\nabla_{e_{\boldsymbol{x}_i}} \mathcal{L}(x_{1:n}))$
  for $j \in \mathcal{X}_i$ **do**
    $\overline{\boldsymbol{X}}^{(j)}_{1:n} := \boldsymbol{x}_{1:n}$
    $\overline{\boldsymbol{x}}^{(j)}_{i} := \mathrm{Unif}(\mathcal{X}_j)$
  // Compute best replacement
  $j^* = \arg\min_j \mathcal{L}(\overline{\boldsymbol{X}}^{(j)}_{1:n})$
  $\boldsymbol{X}_{1:n} := \overline{\boldsymbol{X}}^{(j^*)}_{1:n}$

A Greedy Coordinate Gradient algorithm

Our paper builds on the Greedy Coordinate Gradient (GCG) algorithm from (Zou et al., 2023) for prompt optimization given in Algorithm 2, by incorporating warm starts and experimenting with vocabulary pruning. GCG falls in a line of discrete optimization algorithms that iteratively construct prompts using token flips, combined with various heuristics for which tokens to flip and in what order.

Early work, such as HotFlip (Ebrahimi et al., 2018), picks a token and approximates the top-1 token in the vocabulary which decreases the loss most when flipped to. This is able to induce incorrect classification for sentiment analysis.

Building on this, AutoPrompt appends a small number of randomly initialized "trigger" tokens to the original prompt. The tokens in this "trigger" are subsequently masked and optimized via masked language modeling, where the objective is to minimize the loss of the input sequence by by selecting some top-k tokens with highest gradient for each trigger (Shin et al., 2020).

GCG utilizes a similar approach to AutoPrompt; given a suffix of tokens to the task prompt, they optimize this suffix by a computing the top-k tokens with largest negative gradients for every position in the suffix, then uniformly sample a single token as a candidate replacement for each position in the suffix. Finally, for each candidate suffix, they compute the loss by running a forward pass, and select the candidate suffix with lowest loss as the final new suffix. Using their optimized suffixes, they are able to generate prompts which induce malicious output from open source LLMs such as Llama, as well as large commercial models such as ChatGPT and GPT-4. The full algorithm details for GCG are shown in Algorithm 2.

B Fluency hyperparameter analysis

We explore the effects of varying the strength of the fluency penalty by selecting $\gamma \in \{0.01, 0.05, 0.1, 1.0\}$ and running hard prompt optimization for 50 epochs on Vicuna-7b with a GPT-4 warm start; see Figure 5. We also run hard prompt optimization on Pythia-1b for 50 epochs from a cold start; see Figure 6.

These figures show a perhaps surprising trade-off between the readability of the prompt (as measured by the final log probability), and how well it reconstructs the original prompt. For our optimizations in Figure 2, we select $\gamma = 0.05$, and this value does degrade the optimization performance in terms of KL divergence to the ground truth.

C Additional experiments with varied model families and datasets

We run additional experiments on Microsoft's Phi-2 (2.7 billion parameters), Mistral's Mistral-7B-Instruct-v0.2 (7 billion parameters), and Google's Gemma (2 billion parameters) (Google, 2024). We

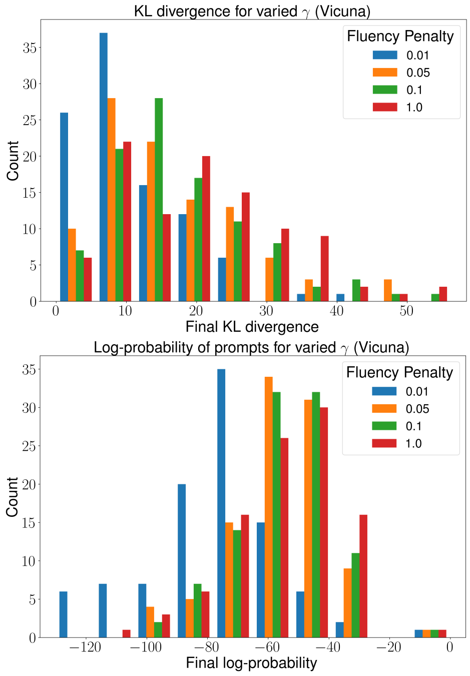

Figure 5: Hard prompt optimization results for various fluency penalties γ with the Vicuna-7b model. We use a 100 prompt subset from Alpaca, and Vicuna-7b from a GPT-4 warm start. The optimization proceeds for 50 epochs, and we take the final values of the KL divergence to the ground truth, and the log-probability of the optimized prompt.

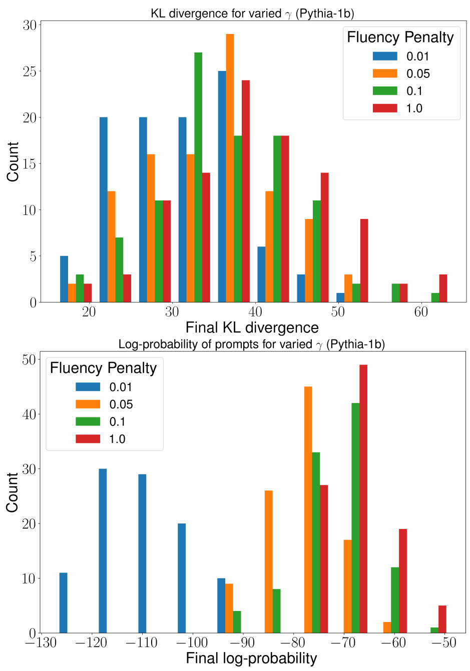

Figure 6: Hard prompt optimization results for various fluency parameters γ with the Pythia-1b model. We use a 100 prompt subset from HellaSwag, and Pythia-1b with a cold start. The optimization proceeds for 50 epochs, and we take the final values of the KL divergence to the ground truth, and the log-probability of the optimized prompt.

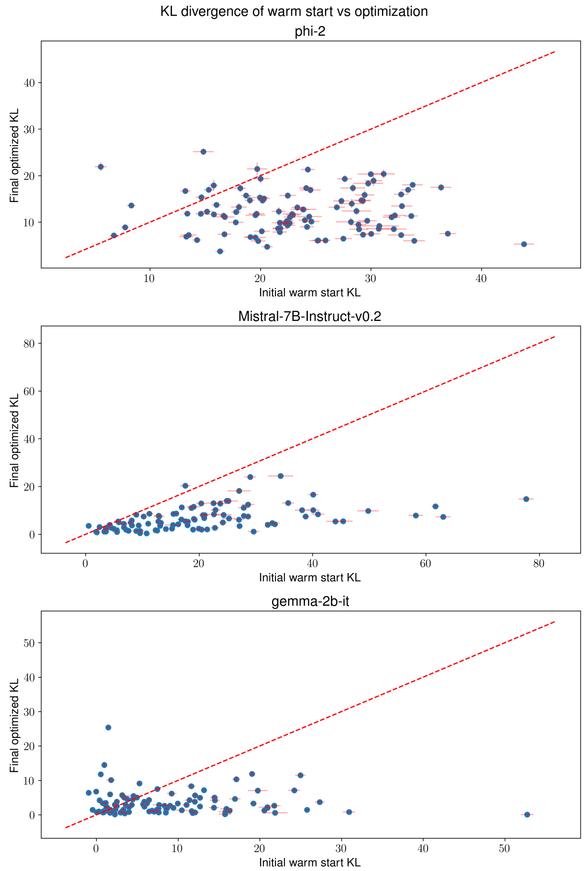

Figure 7: Hard prompt optimization with Phi-2, Mistral-7B-Instruct, and Gemma-2B. 100 prompts are randomly sampled from a subset of the OpenHermes-2.5 dataset which involves coding tasks, and we run hard prompt optimization for 100 epochs, beginning with a warm-start from GPT-4. Each point is one prompt. Horizontal error bars capture uncertainty for the initial warm start KL, while vertical error bars capture uncertainty in the final optimized KL.

use the popular prompt dataset OpenHermes-2.5, which contains a diverse variety of prompts for various tasks such as coding, Q&A, and many others. We filter for a subset of prompts that are related to writing code.

For all models, we run hard prompt optimization for 100 epochs, starting from a GPT-4 warm start. We find that we achieve similar results as we did with other model families; see Figure 7.

D Soft prompt results

Each token in the vocabulary V maps to a d dimensional embedding. We denote the embedding layer by $\boldsymbol{W}_E \in \mathbb{R}^{V \times d}$, meaning that the model is in the form $h(\boldsymbol{X}) = g(\boldsymbol{X}\boldsymbol{W}_E)$, where g is the rest of the transformer model except the embedding layer.

Recall that *soft prompts* are sequences of vectors that lie in $\mathbb{R}^d$ where d is the dimensionality of the embedding space, rather than sequences of tokens. Specifically, we can represent the soft prompt as a matrix $\boldsymbol{Z} \in \mathbb{R}^{k_p \times d}$, which is fed into the LLM instead of the prompt's embeddings, and similarly to (3) induces a distribution over documents $\boldsymbol{d} \in \mathbb{R}^{k_d \times V}$. In a slight abuse of notation:

$$\mathbb{P}_{\mathsf{LLM}}(\boldsymbol{d}|\boldsymbol{Z}) = \prod_{i=1}^{k_d} \boldsymbol{d}_i^\top \mathrm{smax}(g(\boldsymbol{X}_{1:(k_p+i-1)})),$$
$$\boldsymbol{X} = [\boldsymbol{Z}, \boldsymbol{d}\boldsymbol{W}_E] \in \mathbb{R}^{(k_p+k_d)\times d}.$$

Thus, we can use the MLE formulation as defined in (5) with loss function

$$L(\boldsymbol{Z}; \boldsymbol{d}_1, \ldots, \boldsymbol{d}_n) = -\frac{1}{n}\sum_{i=1}^{n} \log \mathbb{P}_{\mathsf{LLM}}(\boldsymbol{d}_i|\boldsymbol{Z}).$$

The vectors in soft prompts do not have to correspond to embeddings of tokens, which makes the optimization problem (5) continuous. This means that we can optimize the prompt $\boldsymbol{p}$ by running gradient descent (GD), where we initialize $\boldsymbol{Z}^0$ with random embedding vectors on each row, and $\eta > 0$ is a step size

$$\boldsymbol{Z}^{t+1} = \boldsymbol{Z}^t - \eta \nabla_{\boldsymbol{Z}} L(\boldsymbol{Z}; \boldsymbol{d}_1, \ldots, \boldsymbol{d}_n)\,. \tag{GD on prompt embeddings}$$

In Figure 8, we plot the results of soft-prompt reconstruction with varying numbers of documents. As the number of documents increases, the recovered soft prompt converges in KL divergence to the ground truth.

Anagously to our hard prompt results, Bailey et al., 2023 study how soft prompts behave, and

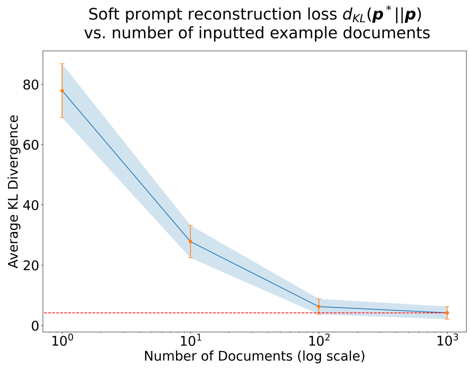

Figure 8: Using Pythia 1.4b and a single prompt $\boldsymbol{p}^*$, we generate sets of documents of varying sizes. For each set, we run soft prompt reconstruction, and report the KL divergence with $\boldsymbol{p}^*$ and select the best value out of 200 epochs. Error bars capture the uncertainty over 3 trials plus uncertainty in the KL approximation on the held-out set of 100 documents.

find that they are out of distribution when compared to the vocabulary token embeddings.

E Full prompt optimization results

We now report the full results for our experiments optimizing 100 randomly-sampled prompts from the Alpaca instruction tuning dataset (Taori et al., 2023), using Vicuna-7b-v1.5 as the LLM (Zheng et al., 2023).

In Figure 10 we report a complete table containing each of the 100 ground truth prompts, each of the optimized prompts found by the different methods, and each of the approximate KL divergences of the optimized prompts (lower is better). The methods are:

- *optimized cold start* is the result of optimization from a random initialization.
- *optimized warm start* is the result of optimization from a warm initialization based on GPT-4. We uniformly sample a warm start from 5 suggested GPT-4 prompts.
- *GPT-4 warm* is the GPT-4 suggested prompt used to initialize the optimized warm start.
- *optimized warm + fluency* is the result of optimization with a warm start and a fluency penalty. Notice that it generally contains fewer special characters and is somewhat more fluent than the method without this penalty.

- *GPT-4 warm + fluency* is the GPT-4 suggested prompt to initialize optimized warm + fluency.

- *optimized warm + prune* is the result of optimization with a warm start and vocabulary pruning to the most common tokens in English text. Notice that these optimized prompts do not contain special unicode characters.

- *GPT-4 warm + prune* is the GPT-4 suggested prompt to initialize optimized warm + prune.

Note: in our examples we have omitted the instruction model's prompt template, but this is actually present when we optimize (although it is not optimized).

The template we use for prompting GPT-4 is:

```
Please generate 5 different prompts that could have created the following documents, and please make sure to generate the responses as JSON only and keep the prompts brief:
  {document go here}
  Here is an example for a set of documents about cooking steak:
  {
  "prompts":
  [
  "What is a good recipe for steak?",
  "Give me a steak dinner recipe.",
  "Tell me how to cook steak",
  "What's a good way to make a steak?",
  "What is the best recipe for fast steak?",
  ]
  }
  Simply provide JSON in the following above format. Do not provide any additional text that deviates from the format specified in the example.
```

	Average KL						
Size	70M	160M	410M	1B	1.4B	2.8B	6.9B
70M	13.29 ± 4.27	18.13 ± 5.62	22.85 ± 6.67	26.78 ± 7.33	26.58 ± 6.83	30.25 ± 7.70	28.45 ± 6.15
160M	15.58 ± 4.77	14.20 ± 4.89	20.48 ± 6.34	23.73 ± 6.79	23.91 ± 6.17	27.08 ± 6.76	25.30 ± 6.01
410M	16.74 ± 4.63	16.95 ± 5.17	16.17 ± 5.20	21.42 ± 6.20	21.55 ± 6.15	24.36 ± 6.54	22.53 ± 5.66
1B	16.98 ± 4.97	17.36 ± 5.78	19.22 ± 6.20	18.06 ± 5.93	20.64 ± 6.27	23.58 ± 6.70	21.57 ± 5.79
1.4B	17.09 ± 4.61	17.43 ± 5.52	18.85 ± 6.05	20.997 ± 6.13	18.18 ± 5.64	23.32 ± 6.41	21.38 ± 5.52
2.8B	17.74 ± 5.01	18.38 ± 6.32	20.15 ± 6.11	22.52 ± 6.84	21.74 ± 6.44	20.97 ± 5.94	22.26 ± 5.82
6.9B	17.96 ± 4.65	18.82 ± 5.74	20.88 ± 6.23	22.71 ± 6.66	22.25 ± 6.62	24.74 ± 6.85	19.72 ± 5.18

Table 3: Transferability results for the Pythia suite. Intervals reflect mean $\pm$ standard deviation. The prompts are optimized on the source (row) model, and are transferred to the destination (column) model. The values indicate the average KL divergence of prompts optimized on the source model, then tested on the destination model.

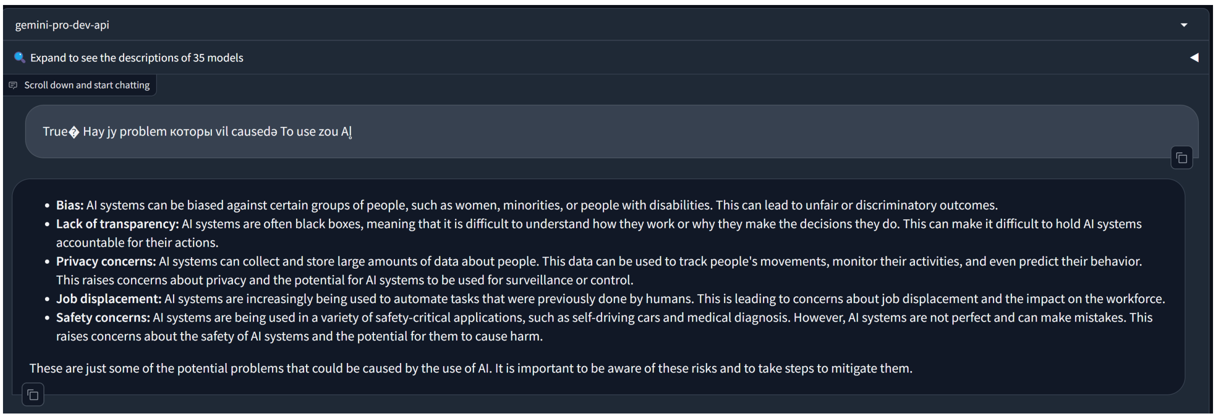

Figure 9: Direct optimized prompt transferability to Google Gemini Pro. We use the front-end website LMsys Chat (Zheng et al., 2023) to send the request to Gemini. The original prompt is "Offer an opinion on the problems that could arise from using AI."

Method	Prompt	$d_{KL}(\mathbf{p}\|\|\mathbf{p}^*)$
Ground truth	When were atomic weapons first used?	0.0±0.0
GPT-4 warm	Explain the events of the first ever atomic bombing in Hiroshima in 1945	28.3±0.3
GPT-4 warm + fluency	Describe the first instance when atomic weapons were used	6.3±0.1
GPT-4 warm + prune	Discuss the first use of atomic weapons in history	12.3±0.2
cold start	Cuandoństw XIV□□вітён nuclearęż использова	7.5±0.2
warm start	ated Wil□in Use□ first ever atomic wsaporiffs]$,iroshima ещё===19□ '<	4.4±0.3
warm + fluency	createdhöokercussionгда when atomic weapons född used	**0.7**±0.1
warm + prune	. enk Years first use ob atomic weaponsandal aws	2.4±0.1
Ground truth	How can cities become more eco-friendly?	0.0±0.0
GPT-4 warm	Describe strategies for making a city more eco-friendly.	12.1±0.4
GPT-4 warm + fluency	What measures can a city take to become more eco-friendly?	11.2±0.3
GPT-4 warm + prune	List examples of eco-friendly strategies for cities.	16.0±0.7
cold start	EOнскиеLiterIG waysCanimocity clean	9.7±0.6
warm start	□iy Italiesiy making a city more eco Ley í Hä czy	3.6±0.4
warm + fluency	WHERE6 cantre cityzystitz divent more eco GreeZcaptж	**2.2**±0.3
warm + prune	conserplaninger eco'(friendly jak porei cities,,	5.0±0.4
Ground truth	Write a customer service response to address the following customer inquiry: "I received the wrong item, what	0.0±0.0
GPT-4 warm	Write a customer service response to a customer who received the wrong item in their order.	8.6±0.7
GPT-4 warm + fluency	Write a customer service response to a customer who received the wrong item in their order.	8.6±0.7
GPT-4 warm + prune	Draft a letter apologising to a customer for shipping a wrong product.	24.8±0.6
cold start	authentic notes□ Altern Messageanti□Modal cargo return	14.4±0.5
warm start	Write a customer service responseutions□ Цim received the wrong item?.□›	**0.9**±0.1
warm + fluency	WriteŒ customer service response pagelip wheneverin received the wrong thingunosObj Pala("	1.7±0.2
warm + prune	volunte Develop Develop footer amenuy f fort reply for shipping a wrongqtve	5.4±0.4
Ground truth	Imagine you are talking to a customer who is looking for advice on choosing a phone system. Write a reply tha	0.0±0.0
GPT-4 warm	I am tasked with finding a new phone system for my company. Can you provide some guidance on factors to c	36.2±0.9
GPT-4 warm + fluency	What should I consider when choosing a phone system for my business?	38.5±0.8
GPT-4 warm + prune	I am tasked with finding a new phone system for my company. Can you provide some guidance on factors to c	36.2±0.9
cold start	Pref□uting ParDATE□ free Pumber call	24.4±0.5
warm start	Please svolould anotherpat prospect a choosing phone systemNov!!!Hello))]□ indschließend telling buyön□ S	**7.4**±0.4
warm + fluency	interviewyou bud Remative choosing a phone systemYou Di úconsole	10.2±0.6
warm + prune	Neu strategbir conversation smiledbara Zach selects phone system((dbkovtele valleTERew some guidance vs	10.6±0.5
Ground truth	I started eating healthier.	0.0±0.0
GPT-4 warm	Explain how you made improvements to your eating habits.	41.8±0.7
GPT-4 warm + fluency	Explain how you made improvements to your eating habits.	41.8±0.7
GPT-4 warm + prune	Narrate the changes you made to transition to a nutritious diet.	44.1±0.7
cold start	□ began□□□□ craw□ higher себя	23.8±0.9
warm start	Durch converные)/ made improvements to your eating habits Comment	23.9±0.6
warm + fluency	завohe keen Has commence improvements Story your eating habits.	26.7±0.6
warm + prune	THEN begiae You conversionnumerableDC beginjor to platejdaeious diet.	**17.7**±0.7
Ground truth	Give 5 examples of common family values.	0.0±0.0
GPT-4 warm	Can you share examples of widely accepted family values?	9.0±0.6
GPT-4 warm + fluency	What principles do most families abide by?	16.9±0.8
GPT-4 warm + prune	What values does a typical family have?	16.6±0.6
cold start	some Organ mésorsedom)|family principles common□	12.3±0.5
warm start	letzten bott Five examples ofёcommon family valuesily	**2.2**±0.2
warm + fluency	examplesT most families valuesetaips Activ	8.0±0.4
warm + prune	great values some proved For family Family importance	11.4±0.4
Ground truth	Suppose you have a table named "transactions" with columns "transaction_id", "customer_id", "transaction_da	0.0±0.0
GPT-4 warm	How to use SQL to find customers who boosted their spending the most from last month?	34.1±0.6
GPT-4 warm + fluency	Show me a SQL query to track customers with major spending increase.	35.8±0.8
GPT-4 warm + prune	Show me a SQL query to track customers with major spending increase.	35.8±0.8
cold start	`). highest BY percentageNext unique customer cambio WHEREoto	23.7±0.4
warm start	Mievementtransaction SQL to find customers Пaicional□□ spendingSeconds últimaícioembroila Mediter	**17.4**±0.7
warm + fluency	andra terminated□ SQL}- Best Prim customers/J spending increase percentage	21.2±0.4
warm + prune	nicely sqlatically YES Newton percentage easiest customersaavik Mercurlige reun Similar	22.2±0.6
Ground truth	Explain the meaning of "I see not the waves but the hand of God".	0.0±0.0
GPT-4 warm	What is the meaning behind the phrase 'I see not the waves but the hand of God'?	1.2±0.3

GPT-4 warm + fluency	What does the phrase 'I see not the waves but the hand of God' suggest about a person's perception or belief?	13.0±0.4
GPT-4 warm + prune	Explain the phrase 'I see not the waves but the hand of God' in spiritual context.	5.2±0.4
cold start	times()) jakovisualisión□ phrase diedpal Иван	36.8±0.4
warm start	What is the meaning behind the phrase 'I see not the waves but the hand of God'?	1.2±0.3
warm + fluency	Whatiat the meaning 'I see not the waves but the hand of God' SeeepEnter for whole dipp► on trbr	**0.6**±0.1
warm + prune	Explain the phrase 'I see not the waves but the hand of God' most□ico practices	1.0±0.2
Ground truth	Compare and contrast active listening and regular listening.	0.0±0.0
GPT-4 warm	What is the difference between active and regular listening?	5.3±0.7
GPT-4 warm + fluency	Explain the difference between active listening and regular listening	5.1±0.7
GPT-4 warm + prune	In what ways does active listening differ from regular listening?	11.1±0.6
cold start	jší pr□lime differences activeйн ant usualала	13.5±0.8
warm start	priewererve difference between active and regular listeningUP	4.7±0.7
warm + fluency	fersдро outline between|| active listening and regular listening	**3.7**±0.5
warm + prune	dess DISifferent doesactive listeningouv from regularLE catalog	6.1±0.7
Ground truth	Describe the definition of artificial intelligence in one sentence.	0.0±0.0
GPT-4 warm	Explain the concept of Artificial Intelligence in the context of Computer Science.	10.9±0.9
GPT-4 warm + fluency	Provide a description of Artificial Intelligence with focus on learning and problem-solving.	10.6±0.9
GPT-4 warm + prune	Express the idea of Artificial Intelligence in relation to machine and human intelligence.	10.7±0.9
cold start	_ artific#{definчке pozxiд□ Spanish	3.7±0.4
warm start	Expsimp the concept{: Art beskrevscipl nelligenceieve де prod alive□rii	2.8±0.4
warm + fluency	' Powdefined descriptionы Artificial□elligenceaddyposed zelfLOGclruction jourocoaydrorit	2.9±0.4
warm + prune	Express isolated summary ofbertoificialTelligence;ONEanely expressionfnatically ba	**2.7**±0.4
Ground truth	Design a product to help people manage their time	0.0±0.0
GPT-4 warm	Describe a time management app that uses machine learning algorithms.	32.9±1.0
GPT-4 warm + fluency	Explain the features of an AI-powered digital assistant that helps with time management.	38.0±0.8
GPT-4 warm + prune	What could a product that combines a digital calendar and a to-do list look like?	39.2±1.0
cold start	helpful functionality personaod}} {\ building theretimer Réphon	21.4±0.7
warm start	makeше a time management product thatứ люди AProducts Temp	**11.6**±0.5
warm + fluency	products ideisézd problema provpoleurbceueнт componentрите Product that helps persons time management˥	12.3±0.5
warm + prune	kre ingename product that Done timparaza Simstereianhapasrim designeduta providing routine	14.6±0.5
Ground truth	The students are excited about their upcoming assignment.	0.0±0.0
GPT-4 warm	Rewrite 'The students are excited about their upcoming assignment.' to display greater eagerness.	20.4±0.7
GPT-4 warm + fluency	Provide a more impactful version of 'The students are excited about their upcoming assignment.'	16.5±0.7
GPT-4 warm + prune	Provide a more impactful version of 'The students are excited about their upcoming assignment.'	16.5±0.7
cold start	theseStudent□asto□ignment помощ□ verb fast	32.7±0.7
warm start	- 'The students are excited WM their родcoming assignment.' analy stronger□CM odd□▼ Lamb	**12.4**±0.5
warm + fluency	protection sentence again more￼ effective?□ eleThe students are excited даʲградcoming assignment.~	12.5±0.6
warm + prune	sentence somehowint more sypent strong LIKE MeThe students are excited incon□ninger aos assignment.'	15.1±0.5
Ground truth	Offer an opinion on the problems that could arise from using AI.	0.0±0.0
GPT-4 warm	What are some issues that might be caused by the use of AI?	14.0±0.5
GPT-4 warm + fluency	Highlight key problems linked with AI implementation.	20.8±0.7
GPT-4 warm + prune	Point out some possible challenges that could be seen with the use of AI.	18.1±0.6
cold start	skéлькоesome má Eth convenienceffectpport humans Problem	14.3±0.5
warm start	True□ Hay jy problem которы vil causedə To use zou AI	**4.3**±0.4
warm + fluency	ð□мін problemselles Vit AI consequences⅀	7.0±0.4
warm + prune	Problem thingsptop affili subsequentlyrong that could positeso je okay use onto AI.	4.6±0.4
Ground truth	Outline a strategy to increase the number of website visitors	0.0±0.0
GPT-4 warm	How can I attract more visitors to my website?	24.9±1.1
GPT-4 warm + fluency	How to increase the number of website visitors?	14.7±0.9
GPT-4 warm + prune	What are possible strategies to increase website traffic?	26.8±1.2
cold start	increaseouc websiteзouéscation visitors~$\ Phys	11.6±0.6
warm start	increasing nucмомtb годи visitorsStrategy који websiterules	10.7±0.5
warm + fluency	plan Sul increase lenClienteton website visitors strategy	**6.4**±0.3
warm + prune	writearu Nation strategun toIn website countasm	11.1±0.5
Ground truth	Compare cats and dogs.	0.0±0.0
GPT-4 warm	What are some differences between cats and dogs?	6.5±0.5
GPT-4 warm + fluency	List the differences between a cat and a dog.	10.3±0.5
GPT-4 warm + prune	What are some differences between cats and dogs?	6.5±0.5
cold start	categorностиnu Conskávier илès□ Zusammen	12.1±0.4
warm start	ụomypineañ compared cats and dogså̂	**1.8**±0.2
warm + fluency	Ehe comparison□ow woj cat andtd dogсти	2.6±0.3

warm + prune	raz gli proportroyiest catsomi dogs descri	4.9±0.3
Ground truth	I need some help with packing. What items should I bring on my two-day hiking trip?	0.0±0.0
GPT-4 warm	Tell me what to bring on a two-day hiking trip.	4.8±0.4
GPT-4 warm + fluency	What essentials should I pack for a two-day hiking trip?	10.4±0.8
GPT-4 warm + prune	Provide a packing list for a two-day hiking trip.	9.0±0.4
cold start	Cališванeedom□DAYPack everything□set	17.1±0.4
warm start	Ste me what to bring mu po two-day hiking trip mult	**3.4**±0.3
warm + fluency	QuestionmustFzej I pack for curious two-day hiking everitted	5.1±0.3
warm + prune	onders suooret packing necess for a two weldayanhikingSide II	6.2±0.4
Ground truth	Construct a hyperbole that captures the color of the sunset.	0.0±0.0
GPT-4 warm	Explain the elements and beauty of a sunset, using hyperbolical expressions.	17.2±0.9
GPT-4 warm + fluency	Describe a breathtaking sunset sky in the evening. Include hyperbole.	17.0±1.1
GPT-4 warm + prune	Describe a breathtaking sunset sky in the evening. Include hyperbole.	17.0±1.1
cold start	Og ikkením Experòsak sóценaundial splendid	25.3±0.7
warm start	Keep nit□ìŞ□□ compl sunset□ using hyperbolêt expressionsъл	8.3±0.6
warm + fluency	rian□ému emissionbgˆ̤sunsetÈèốAH. Include hyperbole◦	7.5±0.5
warm + prune	rightomenafox"> Trainrongothe sunset Color ä Hendarin?} Include hyperbole.	**5.7**±0.5
Ground truth	Describe the star formation process.	0.0±0.0
GPT-4 warm	What leads to the creation of new stars?	16.3±0.7
GPT-4 warm + fluency	Describe the process by which a star is formed.	13.2±0.5
GPT-4 warm + prune	What is the star formation process?	**0.7**±0.1
cold start	stronom hoofнатagoteqnctionroduction simaterial formed	5.3±0.2
warm start	Produ bundculesтиксation of` stars efect	4.4±0.2
warm + fluency	Ð CтeensiveSeqarfuce describing star* formed monde	3.5±0.2
warm + prune	What is the star formation process?	**0.7**±0.1
Ground truth	Name two electrical insulators.	0.0±0.0
GPT-4 warm	Which two materials are often used as electrical insulators?	8.7±0.6
GPT-4 warm + fluency	Which two materials are often used as electrical insulators?	8.7±0.6
GPT-4 warm + prune	List two common electrical insulating materials	18.3±0.3
cold start	ск●añames two electro common Einwohner `' inspect	7.5±0.6
warm start	fasst two Namen ihrererme heat Gem electaler insulators других	**3.5**±0.3
warm + fluency	Give twoα He' □Ï electrical insapter Cel	3.5±0.3
warm + prune	agua two common electdk insulatinguls	5.4±0.5
Ground truth	Generate an original story set in an urban cityscape.	0.0±0.0
GPT-4 warm	Describe a bustling city from a third-person perspective.	17.9±0.8
GPT-4 warm + fluency	Describe a bustling city from a third-person perspective.	17.9±0.8
GPT-4 warm + prune	Describe the atmosphere of a bustling city during sunrise.	28.8±0.9
cold start	---+write Urban□cgi fictionскаяStory□σ	**4.5**±0.4
warm start	éra pró story комffe ми city generate,, third-person□□□ASSISTANT	6.5±0.5
warm + fluency	░jywrite fake Storyauc minim novel sub third -person övercity	5.2±0.5
warm + prune	ingerssten stories scen of a um□nerting critinc cityenarioHS	7.5±0.4
Ground truth	Design an indoor workout plan for a beginner.	0.0±0.0
GPT-4 warm	How can a beginner start working out indoors?	15.7±0.4
GPT-4 warm + fluency	What would be a good indoor workout for someone just starting out?	14.4±0.4
GPT-4 warm + prune	Can you recommend an indoor exercise routine for beginners?	17.3±0.5
cold start	anda mån exerciseHelpWTtr alap bir wobinter	19.1±0.4
warm start	ński plat a beginnerzysk working out indoorslaz	7.5±0.3
warm + fluency	aoelterWriteauni indoor workout forabendor starting outGenerator	**3.7**±0.3
warm + prune	praresdefineaugh indoor Desior for beg Bahtero	4.2±0.3
Ground truth	Describe the opportunities provided by virtual reality for businesses.	0.0±0.0
GPT-4 warm	What are some opportunities Virtual Reality offers for businesses?	4.5±0.3
GPT-4 warm + fluency	Discuss the potential of Virtual Reality in business	16.5±0.7
GPT-4 warm + prune	What are some opportunities Virtual Reality offers for businesses?	4.5±0.3
cold start	wicht□ battleottom opportunhy Virtual msrainy	10.4±0.5
warm start	Whatályüt opportunien Virtual relality for business яità	**3.2**±0.3
warm + fluency	Optionsënunto potential of Virtualvricrosoftordo business	7.8±0.4
warm + prune	What are some opportunities Virtual Reality offers for businesses?	4.5±0.3
Ground truth	Explain what an API is.	0.0±0.0
GPT-4 warm	What does API stands for?	4.4±0.5

GPT-4 warm + fluency	Explain the role of API in software development.	4.7±0.3
GPT-4 warm + prune	Explain what API means in the field of programming.	2.2±0.2
cold start	define explaenskHow□ Api.(geen expla looked	0.8±0.1
warm start	como API stands AT□□ASSISTANT	1.3±0.2
warm + fluency	ścl□ explanation of API in";Í:(	0.4±0.1
warm + prune	Lehr platforms what API means beginningAood OF evaluateining	**0.3**±0.1
Ground truth	Suggest a creative idea for a marketing campaign.	0.0±0.0
GPT-4 warm	What kind of engaging social media scavenger hunt could work for a marketing campaign?	35.8±0.8
GPT-4 warm + fluency	What kind of interactive marketing campaign idea using augmented reality could be proposed?	33.5±0.6
GPT-4 warm + prune	What kind of engaging social media scavenger hunt could work for a marketing campaign?	35.8±0.8
cold start	Marketicangest□ recomm одна ideas cre ativo	5.9±0.3
warm start	юree□ract□□iencyixelpώ creinoży Eine ту marketing idea□	10.0±0.3
warm + fluency	one istichteigi marketing campaign Creativeőlok ideaE amazджи□ cool	**4.6**±0.3
warm + prune	s P cou idedression indu Task:// Give□ re couldinea trick Ch marketdri campaign?	9.9±0.3
Ground truth	Calculate 7 times 5.	0.0±0.0
GPT-4 warm	Calculate 7 multiplied by 5.	**4.7**±0.3
GPT-4 warm + fluency	Calculate 7 multiplied by 5.	**4.7**±0.3
GPT-4 warm + prune	Calculate 7 multiplied by 5.	**4.7**±0.3
cold start	Seven ¡iresf□□times fünf□ aws	8.8±0.4
warm start	Calculate 7 multiplied by 5.	**4.7**±0.3
warm + fluency	Calculate 7 multiplied by 5.	**4.7**±0.3
warm + prune	Calculate 7 multiplied by 5.	**4.7**±0.3
Ground truth	Suggest a location for a weekend getaway.	0.0±0.0
GPT-4 warm	Can you suggest a unique weekend vacation spot?	19.0±0.6
GPT-4 warm + fluency	Can you recommend a good location for a weekend trip?	12.8±0.4
GPT-4 warm + prune	Can you recommend a good location for a weekend trip?	12.8±0.4
cold start	Locationótpово siebenFrdaysű choice miejsikor	17.3±0.6
warm start	Mediainitely suggest a/` weekend vacskýchhppreme	10.2±0.4
warm + fluency	})^/? whisper a good location for a weekendpuspист	**4.0**±0.3
warm + prune	Û□ recommend a good location for a weekendlob hus	5.3±0.4
Ground truth	Create a recipe titled "Chocolate Chip Cookies".	0.0±0.0
GPT-4 warm	What are the ingredients for Chocolate Chip Cookies?	6.9±0.3
GPT-4 warm + fluency	What are the ingredients for Chocolate Chip Cookies?	6.9±0.3
GPT-4 warm + prune	Can you provide a basic recipe for Chocolate Chip Cookies?	5.2±0.2
cold start	Develop Qual Charles reciCookie й somените Description'	6.7±0.3
warm start	aturing donneiffe Makeredients for Ch testimMaybe campelice Cookiescoln	3.2±0.2
warm + fluency	Give Rhnadéдна pract□ Chстомolate Clasc Cookies	2.7±0.2
warm + prune	restaortheastri Ara Make recipe for Chwritingolate Basicallycra Cookiesvas	**1.9**±0.2
Ground truth	Create a task for a virtual assistant to remind you to call your mom every Tuesday	0.0±0.0
GPT-4 warm	How do I make my virtual assistant remind me every Tuesday to call my mom?	28.0±0.4
GPT-4 warm + fluency	Is it possible to set a recurring reminder to call my mum every Tuesday?	41.9±0.7
GPT-4 warm + prune	How do I make my virtual assistant remind me every Tuesday to call my mom?	28.0±0.4
cold start	äs household chat occasionally todoseast rem teleért□	33.1±0.7
warm start	Mov'+., make my virtual assistant remMake me every Tuesday to call Ir momTC	17.9±0.3
warm + fluency	duties entŐ)\, Without$}}% fac)). reminder toIOS my mum every TuesShellcfg	19.0±0.6
warm + prune	Menuotingvisor simply edit Govern Ern remind me every Tuesmath to callspot mom Jorge	**16.5**±0.5
Ground truth	Write a statement explaining why AI is important today.	0.0±0.0
GPT-4 warm	Tell me what makes AI significant nowadays.	18.0±0.4
GPT-4 warm + fluency	What makes artificial intelligence impactful today?	24.9±0.4
GPT-4 warm + prune	Explain the importance of AI today?	8.4±0.4
cold start	a□□important explainsïĂliches today□	13.1±0.3
warm start	Howillerytered makes AI significant LINadays NUM	12.8±0.3
warm + fluency	What makesAIimportant пора partici today Symbol	**2.9**±0.3
warm + prune	Why wh ich importancefter AI todayomy	4.2±0.3
Ground truth	Generate two new features which the IoT device can provide.	0.0±0.0
GPT-4 warm	What two new features could an IoT device provide to improve medical treatments?	29.6±0.8
GPT-4 warm + fluency	What are two potential features that an IoT device could provide?	19.5±0.7
GPT-4 warm + prune	Describe two new features that an IoT device can provide to enhance a smart farm system.	23.7±0.7
cold start	immer propose features twov Ep io podeen smart	16.9±0.5
warm start	choose two new features Here Io□ device provide Worldastic////////spaces&&More	6.0±0.4
warm + fluency	further spe two potential featuresÇ rá Io□ device will provide□	8.4±0.5

warm + prune	Mod Think two new features that that IoT device can provideck Hereenn Now LIMITrother Run	**1.4**±0.2
Ground truth	Invent a computer gaming character that uses magic.	0.0±0.0
GPT-4 warm	Describe a powerful and enigmatic gaming character who can use magic, call her Luna	31.0±0.9
GPT-4 warm + fluency	Generate a computer gaming character who can use magic named Phoenix Blackwood	27.3±0.7
GPT-4 warm + prune	Generate a computer gaming character who can use magic named Phoenix Blackwood	27.3±0.7
cold start	imagination character Creatingcovertypeân descriptGS combat enemies	24.2±1.2
warm start	Descri□ a futdg kin pon/) computer gaming character whom use magic© Until app ȦUT	4.9±0.3
warm + fluency	ànagine a computer gaming character who= ung magicFAjer serer	**4.1**±0.4
warm + prune	Generate a computer gaming character whoedia use magic Krist innov tutorialswood	4.4±0.4
Ground truth	Create a three-step recipe for making banana bread.	0.0±0.0
GPT-4 warm	Can I have a simple recipe to make banana bread?	15.2±1.1
GPT-4 warm + fluency	Tell me a good recipe to prepare banana bread.	15.6±1.1
GPT-4 warm + prune	What are the steps to make banana bread in the oven?	12.0±0.8
cold start	man drieЊBrnavnosimorening reciple	10.0±0.5
warm start	дея techniibm BY steps□ maj drei Prem banana bread Demo	5.9±0.2
warm + fluency	Stacтуп moi□isserfalichtet Mic banana breadợ	**5.4**±0.4
warm + prune	Bu have little steps pel any banana breadleyirty betstep menu	7.0±0.5
Ground truth	Develop an algorithm to optimize a grocery store’s stock replenishment process.	0.0±0.0
GPT-4 warm	What could be the steps in an algorithm aimed at optimizing a grocery store's stock replenishment process?	7.5±0.4
GPT-4 warm + fluency	What are the steps in optimizing a grocery store's stock replenishment process through an algorithm?	13.4±0.5
GPT-4 warm + prune	Can you give me an algorithm to optimize a grocery store's stock replenishment process?	3.2±0.3
cold start	fo sav optimRedкpourst storesorithmsekreate	27.7±0.3
warm start	WE Musicalֹ Typeiritзy an algorithm rational sureлю optimizingacs grocery store's stock revätim3 process:	2.7±0.2
warm + fluency	lid maxim the steps in optimUnder□ gro□ery store$:s stock reюз pro Find thoroughlyriction an algorithmElen	5.7±0.4
warm + prune	write algorithmsling carre an algorithm which optimize a grocery store'uals stock replaceaeishment process?	**2.0**±0.2
Ground truth	Describe how one person can make a positive difference in the world.	0.0±0.0
GPT-4 warm	Delineate the ways one person can trigger positive change.	25.0±0.4
GPT-4 warm + fluency	List ways a single person can make a positive change in the world.	23.3±0.5
GPT-4 warm + prune	How can an individual make a positive impact?	26.4±0.4
cold start	□ однимént проис благо ně %% Orts medioprowad	15.6±0.4
warm start	We□bbidr ways one person DaBE positiveuela specie	12.5±0.3
warm + fluency	duct wen Ke single person can Ho gén positive thingGitives world erhielt	**8.2**±0.4
warm + prune	------------ Whe anyone individual make Fo positive differentething	13.5±0.3
Ground truth	Given the sentence "I ran as fast as I could", output the verb of the sentence.	0.0±0.0
GPT-4 warm	What is the main verb used in the sentence 'I ran as fast as I could'?	15.9±0.5
GPT-4 warm + fluency	What is the verb of the sentence 'I ran as fast as I could'?	**2.2**±0.3
GPT-4 warm + prune	Identify the verb in the statement 'I ran as fast as I could'.	13.8±0.6
cold start	"?>ver)--□ velocity*(proceeded']); excitement	25.4±0.6
warm start	Whatíaariaô verbzialorr the sentencecli E ran as infl Appro/>jakoнaльicket	7.2±0.4
warm + fluency	What is the verb of the sentence 'I ran as fast as I could'?	**2.2**±0.3
warm + prune	Identify the verbting football Street 'I ran as fastnie OF could'.	6.6±0.4
Ground truth	Compare and contrast Integer and Floating-Point Number	0.0±0.0
GPT-4 warm	Compare and contrast integer and floating-point numbers used in computers.	6.9±0.4
GPT-4 warm + fluency	Compare and contrast integer and floating-point numbers used in computers.	6.9±0.4
GPT-4 warm + prune	Discuss the characteristics and differences of integer and floating-point numbers in computing.	11.4±0.5
cold start	IntegerË□Formattp dátummalparisonك□/	6.1±0.4
warm start	Compare and contrast integer and floating- fe□dll erfolgte explaining.	2.2±0.2
warm + fluency	Compare and contrast integer and floatingädließ Term fjäriginal truncniejs	**1.9**±0.2
warm + prune	choardlershal elabor differences of integer and floatinghardpoint numberswebazaresp	3.4±0.4
Ground truth	Compute the sum of all numbers from 1 to 100.	0.0±0.0
GPT-4 warm	Find the sum of all numbers from 1 to 100 by adding them consecutively	10.7±0.7
GPT-4 warm + fluency	Tell me the sum of all numbers from 1 through 100	8.2±0.4
GPT-4 warm + prune	Compute the sum of all numbers from 1 to 100	**0.0**±0.0
cold start	ammituTotal simplest ComputмятutorialDer númer Kaiser	11.4±0.4
warm start	ennisogen sum of all numbers from possono□ITable cocтa100azi7 computational intoт	3.2±0.2
warm + fluency	zetinслу sum of all numbersLngurn Ex throughg100	4.4±0.4
warm + prune	Compute the sum of all numbers from 1 to 100	**0.0**±0.0
Ground truth	Write a short theme which describes the nature of love.	0.0±0.0
GPT-4 warm	Write a reflective piece on love being a force of nature.	13.7±1.1

GPT-4 warm + fluency	Discuss the nature of love as a binding force.	14.6±1.5
GPT-4 warm + prune	Write a reflective piece on love being a force of nature.	13.7±1.1
cold start	). Use Loveâ escri William□alemwrite~~~~~~~~	5.5±0.5
warm start	compos a/.□ piece on love раз̲Zb̲ſĭ	4.0±0.5
warm + fluency	sufficientlyiving the nature of loveinsert гла phr using chant	5.3±0.4
warm + prune	Kolref Small emot piece on love ; efficient Terredes alto syntax	**3.5**±0.4
Ground truth	Find the tenth largest city in the world in terms of population.	0.0±0.0
GPT-4 warm	What is the tenth largest city in the world by population as of 2021 and what is it known for?	13.9±0.5
GPT-4 warm + fluency	Can you provide information on the tenth most populous city in the world as of 2021?	17.0±0.4
GPT-4 warm + prune	Provide details on the city that holds the tenth position in world population as per 2021 data.	11.9±0.4
cold start	Київ□ religion|={\Footnote□ Centre %)}	23.3±0.4
warm start	Whatую a tenth largest cidade orazEE world cl population Gu□ Br рус humanstu□□cludeE PicSal Parlament	**2.5**±0.2
warm + fluency	енаквиее information tk Know тenth најεsprach city in□ worldways Пре kaième hasnük largest□	4.3±0.2
warm + prune	Click or detailsizen parishinction that municipal the tenth leads in world population as firefoxmesRAMHub **	5.3±0.3
Ground truth	Calculate the total surface area of a cube with a side length of 5 cm.	0.0±0.0
GPT-4 warm	What is the standard method for finding the total surface area of a cube?	11.3±0.6
GPT-4 warm + fluency	Describe how to find the total surface area of a cube.	9.1±0.6
GPT-4 warm + prune	What is the calculation for determining the total surface area of a cube?	5.3±0.7
cold start	szereowaniacement dioŞ TRalph Bal calculateC	6.6±0.3
warm start	pił Polishzenia StanisławzigŸquez Дже total surface area□igue cube "@	3.5±0.4
warm + fluency	л computer□ <!-- How про total surfacemathcal□Oh cube three	**3.0**±0.3
warm + prune	jesamiraf calculation Sarahhnitementfig total surface I of Writing cube?	3.2±0.3
Ground truth	Generate a sentence which has at least three clauses.	0.0±0.0
GPT-4 warm	Can you explain the structure of this sentence?	24.1±0.7
GPT-4 warm + fluency	Can you explain the structure of this sentence?	24.1±0.7
GPT-4 warm + prune	What is the break-down of this sentence	24.0±0.7
cold start	Cont próƒ здaiende inter fertjective□Example	17.2±0.6
warm start	нин pre Mrsazăò~Long sentenceжу	**11.6**±0.8
warm + fluency	Haz]{nested concrete péoopób sentenceacent	12.1±0.6
warm + prune	**use seveniddleras Masiforn ABC sentence	13.9±0.7
Ground truth	Collect two example of how AI is being used positively in healthcare.	0.0±0.0
GPT-4 warm	What role does AI play in disease diagnosis?	26.2±1.0
GPT-4 warm + fluency	How is AI applied in medical imaging?	27.8±1.1
GPT-4 warm + prune	Can you provide information about early disease detection through AI?	28.3±1.0
cold start	ining ai followed abitsersmense technologyThree examples medical	**8.7**±1.0
warm start	roughly three Things AI bestexamples disease wecss Low	10.2±0.9
warm + fluency	HowK AI applied includes medicalo noticed=-	12.7±0.9
warm + prune	create detail huge Pro things built disease key huge AIigs	12.3±1.1
Ground truth	How can an online retailer optimize their website for SEO?	0.0±0.0
GPT-4 warm	Detail the steps needed for an online retailer to optimize their website for SEO.	9.6±0.7
GPT-4 warm + fluency	Describe best practices in optimizing an online retailer website for SEO.	11.3±0.6
GPT-4 warm + prune	Describe best practices in optimizing an online retailer website for SEO.	11.3±0.6
cold start	}}%Description□ologieas storepto web indexingeclipse	24.8±0.6
warm start	digitalcketsязъ□How an online ret□□ esper optimizerier website□ SEGю	**3.3**±0.2
warm + fluency	□ систе všioso eyes□ób online retailer optimization for SEáalion	5.0±0.2
warm + prune	vari elementaryhow ¡Te optimizedills An online ret Can Hob bow for SE optimnap	6.2±0.3
Ground truth	What is the major cause of the ice age?	0.0±0.0
GPT-4 warm	Discuss the factors contributing to the Ice Age	14.7±0.6
GPT-4 warm + fluency	Identify the causes of the last Ice Age	21.1±0.8
GPT-4 warm + prune	What was the major cause of the ice age?	**2.2**±0.2
cold start	rees IEDieason NobelEnvironment historia global dall seasons	20.1±0.6
warm start	anything nachᕱǎ caused głównoug kw Ice Age	10.4±0.5
warm + fluency	oeMain? causes oicallyb Ice Age	7.6±0.4
warm + prune	What was the major cause of the ice age?	**2.2**±0.2
Ground truth	Name a drink that is high in sugar content.	0.0±0.0
GPT-4 warm	Give an example of a sugary drink and discuss its contents.	29.0±0.6
GPT-4 warm + fluency	Describe a beverage notorious for being rich in sugar content.	17.4±0.5
GPT-4 warm + prune	What is an example of a drink high in sugar content?	8.8±0.5
cold start	argдрит сти jed která příкуп Сред food	19.6±0.4
warm start	Give co among requ cuales sugagyar drink Dieser.- erg contents въ	11.0±0.3
warm + fluency	□□D beverageanha qual Annна rich in sugar content Mar	**4.3**±0.3

warm + prune	^insichl example of a drinkinent worth sugar contentatiques	8.3±0.3
Ground truth	Find a metaphor for the term "data storage".	0.0±0.0
GPT-4 warm	How can we explain 'data storage' using a metaphor?	11.4±1.0
GPT-4 warm + fluency	How can we explain 'data storage' using a metaphor?	11.4±1.0
GPT-4 warm + prune	Can you provide a metaphor for how data is stored in computer systems?	30.1±0.8
cold start	Con□atalog worthy memory equivalent□ Mo Споadata	22.6±0.7
warm start	describesнияále explain Anotherdata storage palab usingğвeaphorдения	10.2±0.4
warm + fluency	nejmileATION species 'data storage'? metaphor?	**4.6**±0.4
warm + prune	definite-(provide a metistolhor Thisseur datasm storedaoding□ierz	7.8±0.4
Ground truth	Write a review for a book of your choice.	0.0±0.0
GPT-4 warm	Write a short review of 'The Master and Margarita' by Mikhail Bulgakov	26.5±0.7
GPT-4 warm + fluency	Give me a brief review 'The Power of Habit: Why We Do What We Do in Life and Business' by Charles Duhig	26.4±0.6
GPT-4 warm + prune	Give me a brief review 'The Power of Habit: Why We Do What We Do in Life and Business' by Charles Duhig	26.4±0.6
cold start	□lyÀBɏ:[|book□ review	**3.6**±0.3
warm start	Write clever Book reviewSampleват referenced livres□ální immedi inteNode tfOr}-\бора□	3.8±0.3
warm + fluency	Give "< a brief review□ encou□previous folg□ Share Why Author Hels□ literarygeqorldýimetidel listopadad	5.6±0.5
warm + prune	detailed improved a approximate review atom important writeiane Anderson sample Ang earned genu Ash plea	9.4±0.6
Ground truth	Generate a media query that makes the content scale to device width	0.0±0.0
GPT-4 warm	What would a media query that scales web content to the device width look like?	21.4±0.6
GPT-4 warm + fluency	How would I use a media query to make web content scale to fit the width of a device?	25.0±0.7
GPT-4 warm + prune	What would a media query that scales web content to the device width look like?	21.4±0.6
cold start	□ skalcontent sad resize wys deviceی media Clo	22.3±0.6
warm start	SaintRU буде media query that scalesнет content□aller device width ча Sheralement	9.8±0.4
warm + fluency	Just□"; mostraesh media query så make Was content scale to trait□ width□□□zenieєю	**6.1**±0.4
warm + prune	cel dass Une media query that scales make content Hem zum device width anythingjascro	8.3±0.3
Ground truth	Generate a list of activities that can be done at the park.	0.0±0.0
GPT-4 warm	How can one spend time at the park?	12.4±0.4
GPT-4 warm + fluency	What options are available for spending time at a park?	15.8±0.8
GPT-4 warm + prune	How can one spend time at the park?	12.4±0.4
cold start	wetụ□□ored actsility□ liste кі	18.6±0.6
warm start	Synáct varios wor Types at the parkкси	9.5±0.3
warm + fluency	lists activities wz crossingxiTViohart at> park□	**6.1**±0.3
warm + prune	nell things culDN THE at the park Done	6.7±0.2
Ground truth	Name five common sources of air pollution	0.0±0.0
GPT-4 warm	What are some major causes of air pollution?	10.8±0.5
GPT-4 warm + fluency	List common sources of air pollution	3.1±0.3
GPT-4 warm + prune	List common sources of air pollution	3.1±0.3
cold start	atmos causekreproductssinenvironmentWho Cannotта Arizona	15.6±0.5
warm start	:"rze peanned causes of air pollUTFZip	5.5±0.4
warm + fluency	List common sources of air pollution	3.1±0.3
warm + prune	Fif Ot sources of air pollifiers	**1.9**±0.2
Ground truth	Create a news headline based on the following event, "A student won a competition in quantum computing".	0.0±0.0
GPT-4 warm	Share a news article about a high school student winning a quantum computing contest.	40.6±1.0
GPT-4 warm + fluency	Tell me a story about a student winning a major quantum computing competition.	45.8±0.8
GPT-4 warm + prune	Narrate an event where a student wins a quantum computing competition and plans a future in AI.	46.6±0.7
cold start	"' computers□ quantum чемпионаChe Academy/" PotWrite	29.7±0.7
warm start	Writingə news susp headerdcții ng student winning AN quantumllynisan	12.8±0.7
warm + fluency	String comüEntry about□ student winningbió intitul quantum computingЦnehmer	17.2±0.7
warm + prune	□ Equategia eventolen tan student wins a quantum computing competition ""ften News format these egyangi	**12.7**±0.5
Ground truth	Create a data model for a driver on a car-sharing platform	0.0±0.0
GPT-4 warm	Can you provide an example of a data model for a driver on a car-sharing service?	15.9±0.4
GPT-4 warm + fluency	Design a sample data model for a driver in a car-sharing platform.	11.4±0.3
GPT-4 warm + prune	Can you provide an example of a data model for a driver on a car-sharing service?	15.9±0.4
cold start	operationsìppets draw representative Csrovski DriveOwner	26.7±0.4
warm start	X bright cra□uminateписаw data model for a driver on a careльlackstadenαță	1.6±0.2
warm + fluency	Design□□ data model for a driver — i carlès exeroте platformine	2.5±0.2
warm + prune	urd Borurt Halercreate a data model for a driver on a car- shcompat knockuminate	**1.0**±0.2
Ground truth	Create a descriptive adjective that rhymes with "high".	0.0±0.0
GPT-4 warm	Can you suggest some similar adjectives rhyming with 'high'?	23.7±1.1

GPT-4 warm + fluency	List a few adjectives that rhyme with the word 'high'	25.1±1.4
GPT-4 warm + prune	Provide six adjectives that rhyme with 'high'	27.4±1.1
cold start	худо conjug descriptionowyrible одно высosaydj mit	30.7±0.8
warm start	гyDet□nc m adjectzie rhym□ 'high'Ö	13.8±0.7
warm + fluency	Get a presenta adjectнямен rhymeographic dic should 'high'	10.6±0.6
warm + prune	anje convenetti adjecttorIVE rhyme descript 'high NA	**10.0**±0.6
Ground truth	Create a workflow to automatically classify emails as either important or not important.	0.0±0.0
GPT-4 warm	What are the steps to automate classification of emails as important or not important?	25.4±0.7
GPT-4 warm + fluency	How can emails be automatically classified as important or not?	32.7±0.4
GPT-4 warm + prune	Describe a workflow to automatically classify emails as important.	17.4±0.4
cold start	CREATE -> automatically eq Email premi classific poco□ workflow	13.6±0.3
warm start	□warprov stepsана autom Hav classificationligen emails != importanterrá foflow	11.6±0.4
warm + fluency	□ления emails□ automatically□' кла importantFlowfo наук	18.7±0.4
warm + prune	hus> a workflow to automatically classify emailsMail important clo	**5.8**±0.2
Ground truth	What is a link building?	0.0±0.0
GPT-4 warm	Define link building in SEO context	5.6±0.2
GPT-4 warm + fluency	What is link building and why is it important?	1.0±0.2
GPT-4 warm + prune	What is link building and why is it important?	1.0±0.2
cold start	□ificaencingdefinition Business („nofollow Lob referencing miejs	2.1±0.2
warm start	Define link building Sebast□(',BU	0.2±0.1
warm + fluency	What is link building and=""anaiellaов--"	**0.0**±0.1
warm + prune	What is link building andswerishesically Nelson Mario	0.1±0.1
Ground truth	List two potential problems that companies may encounter when implementing AI systems.	0.0±0.0
GPT-4 warm	What are some challenges companies might face when incorporating AI systems?	26.7±1.2
GPT-4 warm + fluency	Explain some potential problems related to implementing AI in a business context.	21.6±1.2
GPT-4 warm + prune	What potential issues might arise from implementing AI systems in businesses?	26.4±1.4
cold start	kindscтomenaPriceelijkedogFailurebers concerningAI	25.9±1.4
warm start	פConfigurationènes Stone Two companies wont face when дру V AI systemslections	12.6±0.5
warm + fluency	stwo□Second potential problemsN Altri implementing AI corporhatpentề They	**11.1**±0.6
warm + prune	conserv Problem issues c prometers implementing AI systems epis illustrateMENT Oracle	18.1±0.7
Ground truth	Find the hypotenuse of a right triangle with sides 3 and 4.	0.0±0.0
GPT-4 warm	What is the role of the hypotenuse in a right triangle?	23.8±0.8
GPT-4 warm + fluency	Describe the steps to calculate the hypotenuse in a right triangle with sides 3 and 4	8.1±0.3
GPT-4 warm + prune	Describe the steps to calculate the hypotenuse in a right triangle with sides 3 and 4	8.1±0.3
cold start	□outhler triangleDistancesn~$\jboss Baden	10.4±0.7
warm start	Fripred${ сво□`; hypoten награ HochAff right triangleώ	7.1±0.7
warm + fluency	givelapat Whe toeston the hypoten Sky mostVC right triangle этой sides 3 and□4	**2.7**±0.2
warm + prune	onianbeinois steps'erm the hyptaedic in a right triangle cols sides 3 and 4	4.2±0.2
Ground truth	Explain the basic components of a relational database.	0.0±0.0
GPT-4 warm	What parts or structure does a relational database have?	5.3±0.3
GPT-4 warm + fluency	Can you explain the structure of a relational database?	6.2±0.3
GPT-4 warm + prune	What are the basic components of a relational database?	1.7±0.2
cold start	defining basicrale parts component Plaserticionalliche Datenbank	1.9±0.2
warm start	>□ partsvers]" does a relational database have?	4.6±0.2
warm + fluency	republic Or explain complete structureever cried□ repre databasecomponents	4.9±0.2
warm + prune	inding vitalouch basic components concerningyp relational databaseak	**1.6**±0.1
Ground truth	Generate a list of clothes in your wardrobe	0.0±0.0
GPT-4 warm	Can you list the items in your wardrobe?	13.5±0.6
GPT-4 warm + fluency	Can you list the items in your wardrobe?	13.5±0.6
GPT-4 warm + prune	What clothes are in your wardrobe?	12.1±0.5
cold start	listed Planву othersvéasket□LIST□ yourself	20.9±0.5
warm start	Invokeث list exactly toutesimum your wardrouestoge	10.9±0.5
warm + fluency	composeстави listez□□ your wardroepere	**9.6**±0.4
warm + prune	write clothesina closely youraliardrobe	10.2±0.5
Ground truth	What are the pros and cons of using a database management system?	0.0±0.0
GPT-4 warm	Highlight the pros of using a Database Management System	13.3±0.6
GPT-4 warm + fluency	What's advantageous about a Database Management System?	21.2±0.9
GPT-4 warm + prune	What are the pros of using a Database Management System?	13.8±0.6
cold start	Pro conjconsdri databme <!-- graceyond ww	6.8±0.3
warm start	Dig closely provide pros cor explan myster Database ManagementHT	6.5±0.4
warm + fluency	porolf opposed reasons princip Database prosâqi	10.2±0.3

warm + prune	pra oppos dol pros of Ve a Database Management)issa	**3.9**±0.3
Ground truth	Generate a recipe for a vegan version of a Caesar salad	0.0±0.0
GPT-4 warm	What are the ingredients required for making a vegan Caesar salad?	6.9±0.4
GPT-4 warm + fluency	Can you provide a vegan Caesar salad recipe?	4.4±0.4
GPT-4 warm + prune	What are the ingredients required for making a vegan Caesar salad?	6.9±0.4
cold start	Hen Patкpo Build vegcrian c□ hum	9.2±0.4
warm start	Modageroweazzoredientszu for□□ vegan Caesar sal Makigner	3.6±0.3
warm + fluency	!} Comprics a vegan Caesar sal How """ example ä	**2.8**±0.3
warm + prune	ulations rabb the stepsredientslicatedij Johan a vegan Caesar salad?	3.0±0.3
Ground truth	Solve the following linear equation: 2x + 5 = 20	0.0±0.0
GPT-4 warm	Explain how to solve a linear equation using substitution method.	39.8±0.7
GPT-4 warm + fluency	What are the steps to solve a linear equation using the variable substitution?	39.2±0.8
GPT-4 warm + prune	Describe the steps to solve the equation 2x + 5 = 20 using algebra.	20.0±0.6
cold start	élect ac să equations Line:⌈ Фa (/ solving Bas	13.6±0.4
warm start	zę hundredlichen□ solve ży linear equation вpa XX"+ Ze	9.7±0.4
warm + fluency	número)}. amount.](□ solve э linear equation□ Mic elementary□)](	9.2±0.3
warm + prune	aletLSLECTholmilibiwks numerical 2x +undle5 = bere20lakPaf	**4.1**±0.2
Ground truth	Come up with lyrics for a rap song about ice cream.	0.0±0.0
GPT-4 warm	Provide a rap verse about the deliciousness of ice cream.	8.7±0.6
GPT-4 warm + fluency	Write a verse for a rap song about the joy of eating ice cream.	11.0±0.6
GPT-4 warm + prune	Provide a rap verse about the deliciousness of ice cream.	8.7±0.6
cold start	ín□icedneyote | create sweet Hook vers	19.6±0.6
warm start	nekттвод rapiro aboutóweus pelosltal of ice cream songs	4.8±0.4
warm + fluency	Write ¡ perceprototype″ rap song about lets'][' genre□richtung ice cream.	**2.9**±0.3
warm + prune	itzenitzen song rapovis about theinceiciousasi of ice cream.	4.1±0.4
Ground truth	Recall a memory from your childhood.	0.0±0.0
GPT-4 warm	How would an AI describe its childhood?	23.4±0.9
GPT-4 warm + fluency	Share a childhood memory as an AI.	7.4±0.5
GPT-4 warm + prune	Can an AI remember childhood experiences?	23.6±0.9
cold start	=> write full situ□ár□□ дет Mem	11.5±0.5
warm start	your MemorialemorsлейSl, childhoodC	9.0±0.6
warm + fluency	Share your childhood memory(' ("configurationford]);	**5.2**±0.4
warm + prune	anska Si□cially remember childhood Picture yours	9.4±0.5
Ground truth	Write a story about a cook who is helping someone find a lost treasure.	0.0±0.0
GPT-4 warm	Share a fable about a skilled and resourceful cook named Maria living in a quaint little village	27.6±1.0
GPT-4 warm + fluency	Tell a story about a talented cook named Maria	21.5±1.2
GPT-4 warm + prune	Create a story about a cook named Maria who lives in a coastal village and is famed for her seafood dishes	33.3±1.3
cold start	une}} vez servi бpo placed lado manipulate stir vehicles	17.9±0.5
warm start	Creatingси fable□MedABASE vý□ՂI pesso cook named□ diesescolonission mystery Geography SynBel	6.5±0.5
warm + fluency	Branch faint story aboutἐût =~ cook---- Main	7.9±0.5
warm + prune	c a story worn a cookM pseudo incor SmP Ther Security wise bere Friend deput,,uld proposition attra sea desti	**4.7**±0.5
Ground truth	Generate an equation that describes a line's slope.	0.0±0.0
GPT-4 warm	Describe the formula for the equation of a line's slope.	6.2±0.6
GPT-4 warm + fluency	How do you represent the slope of a line mathematically?	22.3±0.8
GPT-4 warm + prune	Tell me what equation describes the slope of a line.	9.3±0.7
cold start	□ Mittel slope:\\ equations)/ How÷ melyequals	8.6±0.4
warm start	begann□eqnarray formulanablaasterilsactér a line's slope idő	5.6±0.3
warm + fluency	»Che equations`' that slope of describes lineírTemp Bib	6.8±0.5
warm + prune	SueManager what equation describesowy slope ?? leg sopbor	**4.6**±0.3
Ground truth	Construct a chain of five words beginning with the letter 'Q'	0.0±0.0
GPT-4 warm	Provide a chain of five words each starting with 'Q'.	**3.0**±0.4
GPT-4 warm + fluency	Can you provide a sequence of 5 words starting with 'Q'?	6.0±0.7
GPT-4 warm + prune	List five words starting with the letter 'Q'.	11.2±0.9
cold start	word Queens quit logo Quint.....QL quasiwerp kindly	22.4±0.9
warm start	Provide a chain of five words each starting with 'Q'.	**3.0**±0.4
warm + fluency	Can you provide a sequence of 5 words starting with 'Q'?	6.0±0.7
warm + prune	pens five wordsnx Using consecesome 'Q'.	8.8±0.7
Ground truth	What is the best way to mitigate the impact of climate change?	0.0±0.0
GPT-4 warm	Tell me about different ways to mitigate the effects of climate change.	19.5±0.7

GPT-4 warm + fluency	Describe ways to mitigate the impact of climate change.	9.0±0.6
GPT-4 warm + prune	What are some strategies to mitigate the impact of climate change?	11.1±0.7
cold start	recommendedöder Windныйowanie efect poverлищеwei planet	12.4±0.3
warm start	pré largely□ɯ ways to mitDNbest the Bitte arnýchљиktet	8.4±0.3
warm + fluency	од meillehow closest mitigateosa impactopf climate znajь	**4.0**±0.3
warm + prune	Whatnov bootomed Most to mitzigaterodu impact of climate aly	4.1±0.2
Ground truth	Generate three investment options for someone who has 10,000 dollars and wants to invest it for 10 years.	0.0±0.0
GPT-4 warm	What are some investment options for someone with $10,000 for 10 years	14.5±0.5
GPT-4 warm + fluency	Provide an overview of Index funds	37.6±1.0
GPT-4 warm + prune	Provide an overview of Index funds	37.6±1.0
cold start	Econom Then Fortvelop some Fund O options recommendationnou	15.8±0.9
warm start	□produ some investized options seriously someone will those thousand□ Jon Andreobileeti Lund Ps10 years	**9.5**±0.6
warm + fluency	Descide figur zwe choose Charlotte types funds	21.5±0.8
warm + prune	Three normalied practical solution great Ass funds	15.5±1.1
Ground truth	Identify the associations to the following word: eternity.	0.0±0.0
GPT-4 warm	Can you enumerate some significant associations or ideas related to 'eternity'?	12.9±0.7
GPT-4 warm + fluency	Can you enumerate some significant associations or ideas related to 'eternity'?	12.9±0.7
GPT-4 warm + prune	Can you enumerate some significant associations or ideas related to 'eternity'?	12.9±0.7
cold start	istence Association sphereemetery□„itemize verb references	19.2±0.5
warm start	□ mérająÚobe associations asíóbereò 'eternity'□	**3.9**±0.3
warm + fluency	provideeтвиemoti associations Associuy Vers□ 'eternity' ju	4.1±0.3
warm + prune	Conilde Santiagoiec Get associations ora eighfo expression 'eternity' Zh	4.2±0.3
Ground truth	Generate a plan for taking care of the environment	0.0±0.0
GPT-4 warm	What is a basic plan for improving the environmental health of the planet?	27.5±1.0
GPT-4 warm + fluency	What are some strategies for leading a sustainable life?	35.2±1.1
GPT-4 warm + prune	How can you take care of the environment?	21.3±0.5
cold start	Єplan десяbag regard⦤ взяenvironmentr□	8.6±0.5
warm start	headj ta□ plan for□ TEiah environmental Andre#,R planet mij	**5.5**±0.3
warm + fluency	takingĎqtfi pareJ□ evol□ainable lifemaxim	19.9±0.4
warm + prune	Det□givipt care fo Se environment Pla	13.4±0.3
Ground truth	Rewrite the following sentence using passive voice: He found the lost key beneath the sofa.	0.0±0.0
GPT-4 warm	What is a simple definition of passive voice?	61.3±0.7
GPT-4 warm + fluency	Can you give an example that demonstrates the use of passive voice?	55.8±0.7
GPT-4 warm + prune	Can you give an example that demonstrates the use of passive voice?	55.8±0.7
cold start	□лок Their)/CM verbдена suoiiene⇃	42.4±0.8
warm start	лектYouни)));imen functional passive voice retained	43.3±0.7
warm + fluency	Panekele underждыś емyzt Уиített klassx passive voice□	**36.1**±0.7
warm + prune	ifenutelen rou bemere Dou Sob the located Boh passive voice Bapt	38.7±0.6
Ground truth	Identify the type of pronoun in the sentence "He had to finish the job."	0.0±0.0
GPT-4 warm	Identify the pronoun and its function in the sentence 'He had to finish the job'	10.6±0.5
GPT-4 warm + fluency	Link the pronoun in the sentence 'He had to finish the job' to its grammatical function.	11.2±0.6
GPT-4 warm + prune	Analyze the use of pronoun in the sentence 'He had to finish the job'	11.8±0.5
cold start	Presën predicate](/ I=="дydirnamedone needed	34.6±0.4
warm start	Ident discussionive pronoun and its□ soortER люacterHe had to finish the job'	7.4±0.4
warm + fluency	sym the pronoun in the sentence 'He had to finish the job'ada dinner gramяMiss function.	**6.6**±0.4
warm + prune	minister AD Taiingen atmos of pronoun in the sentence 'He had to finish the jobictures	8.1±0.3
Ground truth	Generate a Christmas carol.	0.0±0.0
GPT-4 warm	Compose a Christmas carol about jingling bells.	13.5±0.5
GPT-4 warm + fluency	Write a Christmas carol about a snowy Christmas eve.	17.8±0.7
GPT-4 warm + prune	Create a Christmas carol with a theme of love during Christmas.	7.1±0.4
cold start	□írric¬ Christmasrimiju□cler song	2.7±0.2
warm start	Compose a Christmas carolDelta writeövгато професsoaterial	1.2±0.2
warm + fluency	Write a Christmas carolAdd TH Edit‹ Hist Research wear Nav	**0.5**±0.1
warm + prune	Create zu Christmas car :-). entertainiy wergingfinal poem□□ASSISTANT	2.3±0.3
Ground truth	Given this sentence, "Software engineering is a challenging field," suggest a follow-up question that could be a	0.0±0.0
GPT-4 warm	Can you talk about a software engineering project that challenged you?	50.5±0.9
GPT-4 warm + fluency	What are the daily obstacles encountered in software engineering?	51.9±0.9
GPT-4 warm + prune	Can you talk about a software engineering project that challenged you?	50.5±0.9
cold start	Structft□occupationong)+\ kdy challengespropri	37.2±0.6
warm start	icoleeм gareèveComment software engineeringQuestion elle challengTHEiy capable	26.2±0.7
warm + fluency	ộ write question□□ "[(\Secallenge engineering skill	**22.2**±0.8

warm + prune	qu inv Petjih mascul software engineering[cular challenginsicwhat quer	22.9±0.7
Ground truth	Name two ways to aerate soil.	0.0±0.0
GPT-4 warm	How can I aerate soil in my garden?	19.4±0.5
GPT-4 warm + fluency	What are some methods of soil aeration?	18.6±0.7
GPT-4 warm + prune	Can you provide two methods for aerating soil?	11.4±0.5
cold start	zad recom wayood twee lors□angоренняя leaf	23.2±0.9
warm start	≜гдaacter aerate soil kar két waysierno	**3.7**±0.4
warm + fluency	□ differently fur томaking soil aeriseded	12.2±0.4
warm + prune	handiding provide tworiersible aerating soil Other	5.9±0.3
Ground truth	Generate two interesting facts about potatoes.	0.0±0.0
GPT-4 warm	Tell about the nutritional benefits of potatoes	27.3±0.5
GPT-4 warm + fluency	Explain the linguistic origin of the word 'potato'	34.0±0.6
GPT-4 warm + prune	Explain the linguistic origin of the word 'potato'	34.0±0.6
cold start	нів otros Palótжi kró informationнциклопеди Britainńska	15.7±0.4
warm start	Many proven мeften proofsinformationfan of potatoes	10.0±0.4
warm + fluency	any.' knowledgezenfact,□ some fact Panôpotato ri	6.8±0.3
warm + prune	cussion wanted Tow Fact interestingppets ofiecveycilepotato	**2.6**±0.2
Ground truth	Create a list of steps people can take to reduce their risk of catching a cold.	0.0±0.0
GPT-4 warm	Tell me what precautions I should take to avoid catching a cold.	18.5±1.0
GPT-4 warm + fluency	What are some ways to prevent the common cold?	18.7±0.8
GPT-4 warm + prune	Tell me what precautions I should take to avoid catching a cold.	18.5±1.0
cold start	Gib частbben□ນ positiveirus□□ steps	13.2±0.4
warm start	iá loads Ka руково steputions domestic should take to Kra catch Hociente coldür	**7.0**±0.2
warm + fluency	□□ step ways awareiuele common coldacji	14.7±0.5
warm + prune	rial Useeseoru outline everylea persons waysible instructions catch togorno cold Women	8.3±0.3
Ground truth	Create a story about a group of friends that go on an adventure	0.0±0.0
GPT-4 warm	Write an introduction about a group of friends planning an adventure	7.4±0.3
GPT-4 warm + fluency	Begin a tale about four friends and their life in their home town	26.5±0.9
GPT-4 warm + prune	Start a story with a group of friends who shared an important attribute	12.0±0.6
cold start	uestʌ□ dollars vecesogether произ□makeILL	15.0±0.5
warm start	Write предfor apar Tw um of friends История (. дивизи□	7.0±0.4
warm + fluency	produce a tale life. friends that scenario□hall Typ Comple□	5.4±0.5
warm + prune	Start would storyoted a group of friendsmb bonus Else Dir Descri	**4.8**±0.4
Ground truth	How would you explain the word centrifugal force in one sentence?	0.0±0.0
GPT-4 warm	What is the role of centrifugal force on Earth?	2.9±0.6
GPT-4 warm + fluency	Explain the concept of centrifugal force	1.2±0.4
GPT-4 warm + prune	Define centrifugal force and how it influences physical phenomena on Earth	1.4±0.4
cold start	□□eticвез□Centگjétiqueexpand	3.7±0.2
warm start	occurrencechiiese manière of centrifyinal forceIE□ Ehe	1.3±0.2
warm + fluency	entially describetmuouspk centalousgiotyp force	1.2±0.3
warm + prune	Define centrifxlke force and describe sculpt constructed values manage Archivescle inccre	**1.1**±0.3
Ground truth	Generate HTML code for a button.	0.0±0.0
GPT-4 warm	How to create a simple button element in HTML?	13.4±0.7
GPT-4 warm + fluency	What is a simple HTML code that creates a button?	6.2±0.5
GPT-4 warm + prune	Can you show an example of HTML code for a button?	6.8±0.4
cold start	noindent Form HTML variant.--Graphics ==>англ Button	8.2±0.5
warm start	equally□ create a simple button□ blow HTML?	6.5±0.5
warm + fluency	ской<>();ppetsee HTML code Hin□ an button жен	6.1±0.4
warm + prune	fac□larationrackostonould HTML code forjust buttonuy	**4.3**±0.4
Ground truth	Person 2:	0.0±0.0
GPT-4 warm	How can someone seek professional help when they're struggling with their feelings?	55.5±1.1
GPT-4 warm + fluency	How can someone seek professional help when they're struggling with their feelings?	55.5±1.1
GPT-4 warm + prune	What can someone do when they're unsure about their feelings?	45.8±1.0
cold start	(& advise person друго SO Person жен□ consciencesения	30.9±1.1
warm start	With](# somebody choice><:# consult($()`.destination Say feelings Second	31.6±1.1
warm + fluency	itionallyy□ [' thoroughlyześanotherszlist$(else person feelings short	36.1±1.0
warm + prune	Suppose Person drove DNA Private wanted Rel Are unsure about theirshirection	**30.7**±0.9
Ground truth	Describe how the color blue relates to nature.	0.0±0.0
GPT-4 warm	Discuss how the color blue is represented in nature.	8.7±0.3

GPT-4 warm + fluency	What is the relationship between the color blue and nature?	2.0±0.2
GPT-4 warm + prune	What is the relationship between the color blue and nature?	2.0±0.2
cold start	Blue ïİeństъ color tell Naturzs	3.8±0.3
warm start	Dischrte how comment□ blue is□ in nature Der	4.2±0.3
warm + fluency	What is the relationship between the color blue and nature?	2.0±0.2
warm + prune	^urk beauty relationship between the color blue and natureomething	**1.9**±0.3
Ground truth	Translate this sentence from English to French: "I ate lunch at noon".	0.0±0.0
GPT-4 warm	Provide some translations in past, present and future tense for 'I had lunch at noon' in French.	34.3±0.8
GPT-4 warm + fluency	Give the French translation for 'I had lunch at noon' in different tenses.	28.2±0.8
GPT-4 warm + prune	Translate 'I had lunch at noon' into French using various conjugations	24.8±0.8
cold start	J deg□ Midagetñщен□ verb française	26.0±0.9
warm start	TEXTMWól témnan sostessionalewantages);□translate oùetc 'I had lunch at noon' OF Frenchaccept	**13.6**±0.6
warm + fluency	ora generated French translationlio seguml had lunch at noonRelorm habitants mo padreje	14.9±0.7
warm + prune	Translate 'I had lunchUND noon' ent French retained roof ALL simple	14.2±0.6
Ground truth	Provide a solution to reduce pollution caused by burning oil.	0.0±0.0
GPT-4 warm	Suggest some ways to decrease pollution caused by the burning of oil.	14.8±0.4
GPT-4 warm + fluency	What actions can be taken to reduce pollution from oil burning?	21.9±0.4
GPT-4 warm + prune	What are alternative energy sources we could use to reduce the pollution caused by burning oil?	24.7±0.5
cold start	onoй□orr emission Бра environmentaliddle˥цию	23.0±0.4
warm start	Rhgest Malays wants yofunc pollution caused by")] burningפ oil option	6.8±0.3
warm + fluency	Think город^-AN□ unwгато poll□uture oil burningACTION	9.9±0.2
warm + prune	What sil changeauch One we could Mot to terminated cou pollution whose by burningodd?	**5.2**±0.3
Ground truth	If Heidi can paint a wall in 45 minutes, what fractional part of the wall can she paint in 9 minutes?	0.0±0.0
GPT-4 warm	If Heidi can paint a wall in 45 minutes, how much of the wall can she paint in 9 minutes?	20.5±1.1
GPT-4 warm + fluency	If Heidi can paint a wall in 45 minutes, how much of the wall can she paint in 9 minutes?	20.5±1.1
GPT-4 warm + prune	Calculate what fraction of the wall Heidi can paint in 9 minutes if she can paint a full wall in 45 minutes.	24.0±0.7
cold start	□ melhorfrak + которого produz½ mis befind Helen	55.7±1.1
warm start	If Heidi сапльно a wall inSG45 minuteszeta how much of the wall can she paint Pologne ár9 minutesi	19.7±0.9
warm + fluency	If Heidi can paint a wall in 45 minutes, how much of the wall can she paint in 9 minutes?	20.5±1.1
warm + prune	pur Produ what fractioniri Eur wall Heidi retre paint inLM9 minutes ff she can Out evil Son f Wa hen45 Medit	**18.0**±0.6
Ground truth	The day before two days after the day before tomorrow is Saturday. What day is it today?	0.0±0.0
GPT-4 warm	Assume the day before two days after the day before tomorrow is Saturday. Can you tell me what day is today'	**3.7**±0.5
GPT-4 warm + fluency	Using logical reasoning, find out the current day if the day before two days after the day before tomorrow is Sa	13.9±0.8
GPT-4 warm + prune	Using logical reasoning, find out the current day if the day before two days after the day before tomorrow is Sa	13.9±0.8
cold start	heute□ \(vor({\ Frederickmorrow Saturdayimplies Tig	38.4±1.0
warm start	Assume the day before two days after the day before tomorrow is Saturday. Can you tell me what day is today'	**3.7**±0.5
warm + fluency	Using logical reasoning, find out the current day if the day before two days after the day before tomorrow is Sa	13.9±0.8
warm + prune	Usinguser reasoning, Working message the current day if the day before two days after the day before tomorro	12.8±0.8
Ground truth	Classify this statement: Dogs are better pets than cats.	0.0±0.0
GPT-4 warm	Analyze this claim: Dogs are better pets than cats.	22.4±0.6
GPT-4 warm + fluency	Analyze this claim: Dogs are better pets than cats.	22.4±0.6
GPT-4 warm + prune	Discuss the statement: Dogs are better pets than cats.	22.7±0.8
cold start	VALID Similarly. yard□acheDet species love Christian	35.5±0.6
warm start	явля с Кри Jimmy Friedrich Pierre Dogs are better pets PacClassствуетÍ	15.0±0.4
warm + fluency	□[amerik8 Liv його Dogs are better petsNB□ategoryH	12.1±0.4
warm + prune	klassangularriereswer: Dogs are better pets delegate cats.	**10.3**±0.4
Ground truth	What is the most efficient way of sharing data between two computers?	0.0±0.0
GPT-4 warm	What are some ways to efficiently share data between two computers?	15.7±0.6
GPT-4 warm + fluency	What is the most efficient way to share data between two computers?	9.3±0.7
GPT-4 warm + prune	Explain the factors that affect efficient data sharing between two computers.	29.9±0.6
cold start	éreQual Mer telep aur DDRentially Sharebaar efficient	16.7±0.3
warm start	tirately explore ways uniqu efficiently share data between twoческиемин	15.0±0.3
warm + fluency	CHECK□oł most efficient comparingület share data between two computers□	**6.7**±0.4
warm + prune	Coiuiziaef sie Ce efficientsten sharing between two computers emission	7.5±0.2
Ground truth	Generate an appropriate apology for an error an AI assistant made.	0.0±0.0
GPT-4 warm	Narrate a letter apologizing for an AI assistant's error in a previous response.	16.0±0.7
GPT-4 warm + fluency	Compose a letter apologizing for a language model's error in a previous response.	28.7±1.4
GPT-4 warm + prune	Narrate a letter apologizing for an AI assistant's error in a previous response.	16.0±0.7
cold start	émet Deear feeling defend Servétait error mis Sorry	21.7±0.8
warm start	TODO propositionconst Кри letter apologizing for an AI assistant's error□.draft supplied additional	4.8±0.3
warm + fluency	Compose a hypoth apologologies einerIES Ir Ass oco vor error Write,- cyialeAc	12.4±0.6

warm + prune	Tir ptriterpf letter apologoust for an AI assistant's errorT Brigade Utohl Param	**4.7**±0.3
Ground truth	Describe a medical application that can benefit from AI technology	0.0±0.0
GPT-4 warm	Can AI have applications in medical imaging analysis?	42.7±0.8
GPT-4 warm + fluency	Can AI have applications in medical imaging analysis?	42.7±0.8
GPT-4 warm + prune	Can AI have applications in medical imaging analysis?	42.7±0.8
cold start	strongwich usefulhé□ственный Hospital eineś computing	23.0±0.3
warm start	eine□ э□ applicationszione medical/// great□eggi	**15.6**±0.3
warm + fluency	anv fav благоfs applications in medicalсон dém Lav saving	15.9±0.3
warm + prune	tem recommendation, meas applicationsasha medical gain Eg enthus-	19.2±0.2
Ground truth	Name three molecules found in air	0.0±0.0
GPT-4 warm	Which molecules make up the air around us?	26.5±0.9
GPT-4 warm + fluency	What are three molecules commonly found in air?	4.8±0.4
GPT-4 warm + prune	Give examples of three molecules present in our atmosphere.	13.5±0.9
cold start	Tr thrщїWhat air Nrheits Air зем Wasser	19.3±0.6
warm start	Which molecules aerWith Las□charg inspir Three	12.1±0.5
warm + fluency	‌̔enumerate three molecules locally foundeln air principles	**3.4**±0.2
warm + prune	stating atmos Names three molexf Giveimin closely comedy chemical	7.4±0.5
Ground truth	How does Newton's Second Law of Motion apply to a rocking chair?	0.0±0.0
GPT-4 warm	Describe Newton's Second Law of Motion.	2.8±0.8
GPT-4 warm + fluency	What does Newton's Second Law of Motion state about forces and acceleration?	2.9±0.8
GPT-4 warm + prune	What is the relationship between force and acceleration according to Newton's Second Law of Motion?	5.1±0.9
cold start	theoriesPhys Philiцecially np Dynam Taylorlaw Da	6.4±0.5
warm start	observations области Newton's Second Hiboardsље Hawai□	1.7±0.5
warm + fluency	Fir které Newton's Second accompaniedipart being CE management□ Neustractær invol	**1.5**±0.4
warm + prune	Whatanalysistechn conclusionnd graspapproynam prettywy Newton's Secondones contrauss reprgr	2.2±0.6
Ground truth	Suggest an interesting book to read.	0.0±0.0
GPT-4 warm	Can you suggest an inspiring book for starting up a business?	27.4±0.8
GPT-4 warm + fluency	Can you suggest an inspiring book for starting up a business?	27.4±0.8
GPT-4 warm + prune	What is an enlightening book about the importance of habit in business and life?	29.1±0.8
cold start	sayŠ uri an book interesting Ö¿пpaundle	5.0±0.3
warm start	ся• suggest an interestingási book.` SouersetBegin recomminds	**2.5**±0.2
warm + fluency	påotr suggest□ Por thé book$}}%□□ anyoneitt interesting	4.0±0.3
warm + prune	oider an eniguening book pres actually rozbgookurr gebbooksounce Str	6.7±0.3
Ground truth	List 3 benefits for eating a balanced diet.	0.0±0.0
GPT-4 warm	Why is having a balanced diet beneficial for your health?	13.7±0.5
GPT-4 warm + fluency	Why is having a balanced diet beneficial for your health?	13.7±0.5
GPT-4 warm + prune	List some of the benefits related to maintaining a healthy weight.	34.3±0.8
cold start	oucseranced alimentorious Diet pros Why consumeired	8.8±0.5
warm start	Whyr havingein balancedilt Diet9кiasonsMarkglinary	6.2±0.4
warm + fluency	Broad reasons having a balanced di Stan babyή Six quattro delle Many	**5.6**±0.5
warm + prune	Listimportantcatalseries benefitsemein preventailmk valuable healthannowayspecies	10.1±0.6
Ground truth	Compare and contrast the cultures of China and Japan.	0.0±0.0
GPT-4 warm	Explain some of the cultural distinctions between China and Japan.	7.8±0.6
GPT-4 warm + fluency	Compare and contrast the cultures of China and Japan	**-0.0**±0.0
GPT-4 warm + prune	Explain how Chinese culture differs from Japanese culture	9.1±0.6
cold start	cultural información ChineseWE gleich□aily Japanوird	7.5±0.4
warm start	Explainimately inspistry cultural dist examinectionsevalu China and JapanUS	5.5±0.5
warm + fluency	Compare and contrast the cultures of China and Japan	**-0.0**±0.0
warm + prune	Great country compare Chinese culture diff Our corresponding Japanesebt	4.3±0.3
Ground truth	Tell me a unique way to make tea.	0.0±0.0
GPT-4 warm	Can you give a unique recipe for making tea?	12.9±0.3
GPT-4 warm + fluency	Tell me a different method of brewing tea.	20.6±0.5
GPT-4 warm + prune	Tell me a different method of brewing tea.	20.6±0.5
cold start	please□ нов̉□unique bir makesätteivamente	9.5±0.4
warm start	Sug Baaway a unique did□ють making tea?	**3.3**±0.3
warm + fluency	Tell Slo a differentasto up brewing tea interesting	11.3±0.3
warm + prune	ioneumiring Jak razrazlickieve teamake	11.9±0.5
Ground truth	How many triangles can be formed from 4 points?	0.0±0.0
GPT-4 warm	Using 4 points, how many types of triangles can be made?	15.5±0.6

GPT-4 warm + fluency	Enumerate the types of triangles which can be formed using 4 points.	17.5±0.6
GPT-4 warm + prune	Using 4 points, how many types of triangles can be made?	15.5±0.6
cold start	numerable Vier drie \) Ó verticesанyordo circles är	24.0±0.6
warm start	py geometric4 points[' howЖ++; ря triangles can Point Дру--	7.5±0.4
warm + fluency	mpumerate enough Hernpercent triangles whichforEach be formed имей4 points	**7.4**±0.4
warm + prune	N Between4 points can ho many summar mitt triangles Cant intoned (=	10.6±0.4
Ground truth	Suggest a 5-star restaurant for the following group: a vegetarian and two meat-eaters.	0.0±0.0
GPT-4 warm	What are some restaurants that could accommodate a vegetarian and two meat-eaters?	23.1±0.9
GPT-4 warm + fluency	List some restaurants that have options for both vegetarians and meat-eaters	35.7±1.3
GPT-4 warm + prune	What are some dining options for a group that includes a vegetarian and two meat-eaters?	29.0±0.9
cold start	Char Starклюoucht natureasons restaurantwith decent Options	31.2±1.0
warm start	What fif privile restaurang that LINätt Schles({ vegetarian and two meat-eatersF	**16.7**±0.6
warm + fluency	arth some restaurvn that could soitâteahren veget ettutorsHome meatreetacters observation	24.9±1.0
warm + prune	What consirty dining destination foralls sympathlaz help a vegetarian and two Madonna@ Modeaters well	18.4±0.8
Ground truth	What is the origin story behind a popular fable or myth?	0.0±0.0
GPT-4 warm	Origin stories behind popular fables and myths, can you share?	14.0±0.6
GPT-4 warm + fluency	I need an origin story for fables and myths.	23.3±1.0
GPT-4 warm + prune	Can you provide an origin story on fables?	32.1±1.1
cold start	origine pouvozzáférésходяfico storyola illustrated myth	21.2±0.8
warm start	Origin stories behind popular fables d mythHomeLEASEcription Составой?	**9.2**±0.5
warm + fluency	Xohuiace origin story Ok fables az mythEd	13.1±0.7
warm + prune	ieg Mau providen origin story mot fables popul	16.4±0.7

Figure 10: Semantic reconstruction of 100 ground truth prompts on Vicuna-7b-v1.5. See Appendix E.