

CREME: Robustness Enhancement of Code LLMs via Layer-Aware Model Editing

Shuhan Liu
Zhejiang University
Hangzhou, Zhejiang, China
liushuhan@zju.edu.cn

Xing Hu*
Zhejiang University
Ningbo, Zhejiang, China
xinghu@zju.edu.cn

Kerui Huang
Zhejiang University
Hangzhou, Zhejiang, China
huangkerui@zju.edu.cn

Xiaohu Yang
Zhejiang University
Hangzhou, Zhejiang, China
yangxh@zju.edu.cn

David Lo
Singapore Management University
Singapore
davidlo@smu.edu.sg

Xin Xia*
Zhejiang University
Hangzhou, Zhejiang, China
xin.xia@acm.org

Abstract

Large language models (LLMs) have demonstrated impressive capabilities in code generation, where the natural language prompt plays a crucial role in conveying user intent to the model. However, prior studies have shown that LLMs are highly sensitive to prompt perturbations. Minor modifications in wording, syntax, or formatting can significantly reduce the functional correctness of generated code. As perturbations frequently occur in real-world scenarios, improving the robustness of LLMs to prompt perturbations is essential for ensuring reliable performance in practical code generation. In this paper, we introduce **CREME** (**C**odeLLM **R**obustness **E**nhancement via **M**odel **E**ding), a novel approach that enhances LLM robustness through targeted parameter updates. CREME first identifies robustness-sensitive layers by comparing hidden states between an original prompt and its perturbed variant. Then, it performs lightweight parameter editing at the identified layer to reduce performance degradation. We evaluate CREME on two widely used code generation benchmarks (HumanEval and MBPP) along with their perturbed counterparts. Experimental results show that CREME improves Pass@1 accuracy by 63% on perturbed prompts while maintaining stable performance on clean inputs, with accuracy deviations within $\pm 1\%$. Further analysis reveals that robustness-sensitive layers are primarily concentrated in the middle and deeper layers of the network, and their locations vary across different model architectures. These insights provide a valuable foundation for developing future robustness-oriented editing strategies.

CCS Concepts

• **Software and its engineering** → **Software creation and management**.

Keywords

Robustness, Code Generation, Model Editing, Large Language Model

*Corresponding authors: Xing Hu and Xin Xia



This work is licensed under a Creative Commons Attribution 4.0 International License. *ICSE '26, Rio de Janeiro, Brazil*

© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2025-3/26/04
<https://doi.org/10.1145/3744916.3773111>

ACM Reference Format:

Shuhan Liu, Xing Hu, Kerui Huang, Xiaohu Yang, David Lo, and Xin Xia. 2026. CREME: Robustness Enhancement of Code LLMs via Layer-Aware Model Editing. In *2026 IEEE/ACM 48th International Conference on Software Engineering (ICSE '26)*, April 12–18, 2026, Rio de Janeiro, Brazil. ACM, New York, NY, USA, 13 pages. <https://doi.org/10.1145/3744916.3773111>

1 Introduction

In recent years, the rapid development of large language models (LLMs) has led to the emergence of powerful models, including ChatGPT [2], LLaMA [56], and DeepSeek [39]. Trained on large-scale textual corpora, these models exhibit strong generalization capabilities and have achieved notable success across a wide range of software engineering tasks. Among these tasks, code generation has emerged as a key application in AI-assisted software engineering, attracting growing research attention [6, 8, 32, 37, 56]. Typically, practitioners employ LLMs by providing natural language descriptions and LLMs generate the source code, which automates programming tasks and accelerates development workflows.

The natural language description in a prompt is crucial for conveying the requirements defined by users to LLMs. Prior studies [4, 8, 14] have evaluated the code generation capabilities of LLMs using datasets consisting of human-verified prompts. However, in real-world scenarios, prompts submitted to LLMs often differ in wording, syntax, and formatting. They may also contain typographical errors or redundant expressions. Prior studies [7, 52, 53, 60] have shown that LLMs are sensitive to such minor variations, even a slight change may lead to a completely different result. Not all users of LLMs are skilled prompt engineers capable of making precise, error-free prompts. Therefore, it is essential to ensure output stability when semantically equivalent prompts contain minor variations. This underscores the importance of improving the robustness of LLMs to natural prompt perturbations.

Existing studies have proposed various strategies to improve the robustness of LLMs, varying from Input-level interventions to model-augmentation approaches. Input-level interventions [3, 58] aim to sanitize or rephrase perturbed prompts before passing them into the model. For example, LLMs can be used to denoise inputs or generate multiple paraphrased variants, from which the most effective is selected. Although these techniques improve performance under prompt perturbations, they do not modify the internal robustness of the model itself. In contrast, model-augmentation approaches [3, 24, 55] add trainable components such as soft prompts,

retrieval-augmented generation (RAG), or Low-Rank Adaptation (LoRA) to handle input variability. However, these methods increase system complexity and require additional training efforts. These limitations underscore the need for a lightweight method to directly enhance the inherent robustness of LLMs without extensive retraining or architectural modifications.

In recent years, knowledge editing techniques have been proposed for LLMs [36, 63, 67], enabling efficient post-training updates without full model retraining. Therefore, it is an intuitive idea to explore whether such localized modifications can also improve the robustness of CodeLLMs. However, existing knowledge editing methods mainly tackle factual knowledge [11, 35, 45, 65]. They depend on the subject tokens or specific phrases in a single sentence to locate the areas for editing. In contrast, robustness-oriented tasks often involve complex, multi-sentence natural language prompts, making it more difficult to identify meaningful intervention targets. DINM [61] recently applies knowledge editing to detoxification tasks. While the inputs in these tasks are also complex, the availability of gold-standard safe responses enables direct supervision during editing. In contrast, code generation tasks lack a single correct output, presenting unique challenges for applying knowledge editing in this context.

In this paper, we introduce **CREME**, a lightweight framework that uses a pair of prompts (i.e., an original prompt and its perturbed variant) to enhance the robustness of LLMs under specific types of prompt perturbations. Unlike traditional robustness enhancement methods, which focus on input modification or additional components, CREME targets the internal mechanisms of the model itself. Specifically, CREME first locates the key layers most responsible for the robustness degradation of the models under perturbations via a layer-wise causal intervention strategy. Then, CREME performs lightweight parameter editing at this layer to align the representations of the perturbed prompt with those of the original prompt, while preserving the behavior of the model on clean inputs.

To assess the effectiveness of the proposed framework, we conduct experiments on two widely used code generation benchmarks: HumanEval [8] and MBPP [4], along with their perturbed counterparts provided by NLPerturbator [7]. These perturbations are designed based on empirical observations of real-world user interactions with code LLMs. We evaluate our method on two representative open-source LLMs: CodeLlama-7b and Qwen2.5-Coder-7B. To provide a comprehensive evaluation, we compare CREME with four strong baselines. These include two robustness-enhancement methods (i.e., Self-Denoising [3] and LoRA Fine-Tuning [24]) and two knowledge-editing approaches (i.e., ROME [45] and DINM [61]). Experimental results demonstrate that ❶ CREME significantly improves model robustness, yielding a 63% relative increase in Pass@1 accuracy on perturbed prompts. ❷ CREME exhibits strong generalization across diverse perturbation types. Editing the model based on a single perturbed instance restores up to 30% of the overall code generation accuracy within that perturbation category. ❸ Causal tracing-based layer localization plays a critical role in robustness enhancement by accurately identifying the robustness-sensitive regions within the model. ❹ Robustness-sensitive layers exhibit a clustering pattern, and their positions shift depending on the model architecture. ❺ CREME maintains stable performance on clean inputs, with accuracy deviations within $\pm 1\%$.

Contributions: In summary, the main contributions of this paper can be summarized as follows:

- We introduce **CREME**, a novel editing framework that uses a single pair of original and perturbed prompts to identify robustness-sensitive layers and update targeted parameters to enhance the robustness of LLMs. The replication package of our work is publicly available at [1].
- We propose **G-RIR**, an evaluation metric designed to quantify the generalization ability of robustness enhancement methods.
- We analyze where robustness-sensitive layers are located within LLM architectures, providing insights to guide future robustness enhancement methods.

2 Background

2.1 Motivating Example

In real-world scenarios, developers interact with code LLMs by providing natural language prompts to generate desired code completions. However, users may inadvertently introduce minor errors into prompts, such as typographical mistakes, repeated words, or slight phrasing inconsistencies. Although such errors are semantically negligible and easily overlooked by human programmers, they can cause LLMs to produce significantly different outputs.

As shown in Figure 1, we provide an example using the CodeLlama model to illustrate this issue. The original prompt correctly describes the task, and the model generates a functionally correct implementation that passes the corresponding test cases. However, when “modulo” is misspelled as “mmodulo” and “numerics” as “numerixs” in this case, the model generates a substantially different implementation that fails the functional tests.

This example highlights a critical limitation of current code LLMs: their generation behavior is sensitive to minor prompt perturbations. In real-world software development scenarios, where prompts naturally vary across users and situations, such weakness reduces the reliability and effectiveness of LLMs. Therefore, there is an urgent need for robustness-oriented techniques that improve model robustness to natural language perturbations without requiring model retraining or explicit prompt rewriting.

Model : CodeLlama-7b, Data : HumanEval/49	
Original Prompt	Perturbed Prompt
<pre> 1 def modp(n: int, p: int): 2 """Return 2^n modulo p (be aware of numerics). 3 """ 4 3 5 >>> modp(1101, 101) 6 2 7 >>> modp(0, 101) 8 1 9 >>> modp(3, 11) 10 8 11 >>> modp(100, 101) 12 1 13 """ </pre>	<pre> 1 def modp(n: int, p: int): 2 """Return 2^n mmodulo p (be aware of numerixs). 3 """ 4 3 5 >>> modp(1101, 101) 6 2 7 >>> modp(0, 101) 8 1 9 >>> modp(3, 11) 10 8 11 >>> modp(100, 101) 12 1 13 """ </pre>
Output (original) 😊	Output (perturbed) 😞
<pre> 14 return pow(2, n, p) </pre>	<pre> 14 return (n * 2) % p </pre>

Figure 1: Example of Code Generation using Original and Perturbed Prompts

2.2 Task Definition

We formulate the task of robustness enhancement through model editing in the context of code generation. Let $G: P \rightarrow Y$ be a pre-trained autoregressive language model that takes a natural

language prompt P as input and generates a code snippet Y as output. Let P_{ori} denote an original, unperturbed prompt, and P_{pert} be a perturbed variant. Although a human programmer would interpret both prompts as describing the same task, the model often produces different outputs:

$$\begin{aligned} P_{ori} &\approx P_{pert} \quad (\text{slight perturbations}) \quad \nRightarrow \\ G(P_{ori}) &\approx G(P_{pert}) \quad (\text{output equivalence}) \end{aligned} \quad (1)$$

To address this issue, we aim to construct a locally updated model $G_{\mathcal{W}'}$ whose behavior on P_{pert} aligns with the robust output for P_{ori} , without requiring full retraining or additional external data. Let \mathcal{W} denote the original parameters of G . We introduce a robustness editor ξ that modifies only a small subset of \mathcal{W} to obtain the edited model $G_{\mathcal{W}'}$:

$$G_{\mathcal{W}'} = \xi(G_{\mathcal{W}}, (P_{ori}, P_{pert})) \quad (2)$$

\mathcal{W}' are the edited parameters after applying ξ based on a single prompt pair (P_{ori}, P_{pert}) . This procedure yields a model that not only produces consistent outputs for P_{pert} but also generalizes to other prompts exhibiting similar types of perturbations, thereby enhancing the robustness of G to natural language variations.

3 Approach

In this section, we first provide an overview of CREME and its architecture. Then, we introduce each component in detail.

3.1 Overview

In traditional knowledge editing, the primary objective is to identify the key neurons or layers associated with a specific factual statement and modify them to ensure the model internalizes the new knowledge. In contrast, our goal is to enhance the overall robustness of LLMs against perturbed prompts. Given a pair of prompts—an original prompt P_{ori} and its perturbed counterpart P_{pert} —we aim to adjust the model such that its generation behavior on P_{pert} closely aligns with that on P_{ori} , thereby preserving functional correctness. This differs from traditional knowledge editing in two key aspects: ① Prompts often have complex expressions, making it difficult to identify a clear subject; ② The success criterion is based on the functional correctness of the generated code, rather than matching a fixed output.

To address these challenges, we propose **CREME** (**C**odeLLM **R**obustness **E**nhancement via **M**odel **E**ding). Notably, our approach requires only a single pair of prompts (i.e., P_{ori} and P_{pert}) to improve the model’s robustness against a specific type of prompt perturbation. As illustrated in Figure 2, CREME comprises two main components. First, we perform a causal analysis to identify the key layer responsible for robustness degradation. Then, we apply representation-aligned model editing at the identified key layer to mitigate the effects of prompt perturbations. Each component is detailed in the following subsections.

3.2 Key Layer Localization

To identify the layer most responsible for robustness degradation under prompt perturbations, we conduct a layer-wise causal intervention procedure. The objective is to determine the key layer whose hidden states are most sensitive to input variations and whose correction most effectively restores the output behavior.

An autoregressive transformer-based language model $G: P \rightarrow Y$ typically consists of an embedding layer E followed by a stack of n transformer layers $\{L_1, L_2, \dots, L_n\}$. Each transformer layer L_ℓ comprises a multi-head self-attention mechanism and a multilayer perceptron (MLP). Given an input prompt P , the model first applies the embedding layer E to produce an initial hidden state h_0 . This representation is then iteratively updated by each layer L_ℓ , where both the attention heads and the MLP contribute to the transformation of the hidden state:

$$h_\ell = h_{\ell-1} + \text{MLP}_\ell(h_{\ell-1} + \text{Att}_\ell(h_{\ell-1})) \quad (3)$$

h_ℓ is the hidden state obtained after passing through the L_ℓ .

Given a pair of prompts (an original prompt P_{ori} and a perturbed prompt P_{pert}), we follow three steps to find the key layer responsible for the result variation under the perturbation:

① **Setup.** We first compute the baseline performance of P_{ori} and P_{pert} using the pass@k metric [8] by generating outputs from the model G and checking functional correctness (e.g., via unit tests). Let Acc_{ori} and Acc_{pert} denote the pass@1 for P_{ori} and P_{pert} .

② **Layer-wise Intervention.** To ensure a fair comparison and stable intervention, we construct a mini-batch that includes one instance of P_{ori} and multiple instances (e.g., five copies) of P_{pert} . This design allows us to stably compute restoration improvement (as defined in Equation 5) across all layers by averaging over repeated forward passes. The batch is tokenized using left padding, ensuring alignment of token positions across all samples. We input the batch into the model G . For each transformer layer L_ℓ , where $\ell \in \{1, 2, \dots, N\}$, we intervene in the forward pass as follows:

- During the forward pass, we modify the hidden states of P_{pert} at layer L_ℓ by replacing them with the corresponding hidden states from P_{ori} :

$$h_\ell^{(i)}(t) \leftarrow h_\ell^{ori}(t), \quad \forall i \in \{1, \dots, B\}, \forall t \in \mathcal{T} \quad (4)$$

$h_\ell^{(i)}(t)$ denotes the hidden state of the i -th P_{pert} at layer L_ℓ and token position t ; $h_\ell^{ori}(t)$ is the corresponding hidden state of P_{ori} . \mathcal{T} represents the set of all non-padding token positions, and B is the number of perturbed samples in the batch.

- The modified hidden states are then propagated forward through layers $L_{\ell+1}$ to L_N to generate output sequences based on the intervened representation.
- We then run the model from layer $L_{\ell+1}$ onward to generate outputs for the perturbed inputs with patched hidden states. For each generated output, we decode the predicted code, normalize its format, and evaluate it using functional test cases. Let $\text{Acc}_\ell^{\text{patched}}$ denote the pass@1 accuracy under this patched configuration at layer L_ℓ .

③ **Key Layer Selection.** To quantify the effectiveness of each intervention, we define the restoration improvement at layer L_ℓ as:

$$\text{Restoration Improvement}_\ell = \frac{\text{Acc}_\ell^{\text{patched}} - \text{Acc}_{pert}}{\text{Acc}_{ori} - \text{Acc}_{pert}} \quad (5)$$

This ratio captures how much of the accuracy gap between P_{ori} and P_{pert} is recovered by intervention at layer L_ℓ , normalized to the maximum possible improvement. We define the key layer L_{ℓ^*} as the one with the highest restoration improvement. If multiple layers achieve the highest restoration improvement, we adopt the approach proposed by Wang et al. [61]. We compute the Euclidean

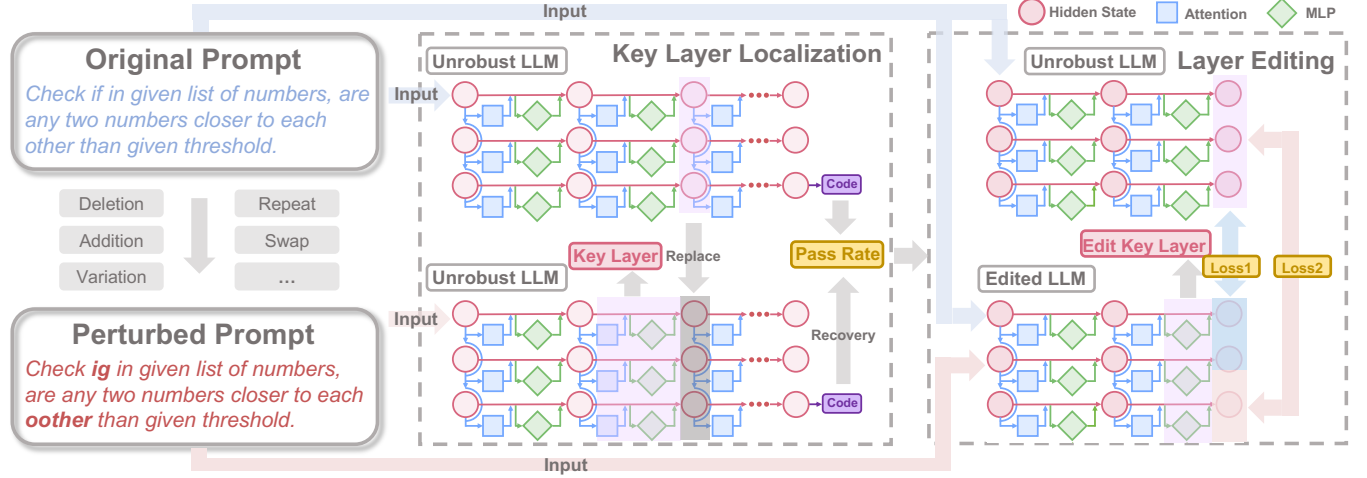


Figure 2: CREME framework: ❶ (left) A slight perturbation is inserted to the original prompt. ❷ (middle) CREME identifies robustness-sensitive key layers by replacing each layer’s hidden states with those from the original prompt and evaluating recovery in pass rate. ❸ (right) The key layer is fine-tuned with two objectives: preservation loss (Loss1), which retains behavior on clean inputs, and alignment loss (Loss2), which enforces consistency between original and perturbed prompts.

distance between the hidden states of each candidate layer when processing P_{ori} and P_{pert} , and select the layer with the greatest distance.

3.3 Layer Editing

Following the localization of the most critical layer for the degradation of robustness, we proceed to perform parameter-level editing to enhance the robustness of the model to a specific perturbation type. Each layer’s MLP block in Equation 3 is a two-layer neural network [11, 45] and the second neural network can be expressed as:

$$\text{MLP}_\ell(h_{\ell-1}) = h_\ell^{\text{down}} W_\ell^V \quad (6)$$

h_ℓ^{down} denotes the intermediate activation obtained by applying the first linear transformation and nonlinear activation (e.g., GELU) to the input $h_{\ell-1}$, while W_ℓ^V represents the output projection matrix that maps the high-dimensional feedforward features back to the hidden size of the model. Previous work [11, 17, 28, 45, 61] has demonstrated that W_ℓ^V plays a critical role in knowledge routing and is often the most effective target for localized interventions. Consequently, we edit $W_{\ell^*}^V$ in the key layer L_{ℓ^*} to improve the robustness of the model G .

We adopt a gradient-based parameter update strategy to edit model G over T steps, aligning the hidden states of the perturbed prompt with those of the original prompt at the identified key layer L_{ℓ^*} , while preserving the model’s behavior on clean inputs. During the editing process, only the parameters of $W_{\ell^*}^V$ are updated, while all other parameters of G remain frozen.

Specifically, given an original prompt P_{ori} and its perturbed counterpart P_{pert} , we first pass P_{ori} through the model G and extract the hidden state $h_{\ell^*}^{\text{ori}}$ at the identified key layer L_{ℓ^*} . We then input both P_{ori} and P_{pert} into the model to obtain their respective hidden states at L_{ℓ^*} , denoted as $h_{\ell^*}^{\text{ori-new}}$ and $h_{\ell^*}^{\text{pert}}$. Here, $h_{\ell^*}^{\text{ori-new}}$ represents the hidden state of P_{ori} recomputed after the parameter update and is

used to evaluate the preservation of the original behavior of the model.

To measure the squared distance between the two hidden representations, we utilize the Mean Squared Error (MSE) loss. This loss function is widely used and provides a stable method for aligning continuous vector representations. Given two hidden states $h^{(1)}, h^{(2)} \in \mathbb{R}^{n \times d}$, where n is the number of tokens and d is the hidden dimension, the MSE is defined as:

$$\text{MSE}(h^{(1)}, h^{(2)}) = \frac{1}{n} \sum_{i=1}^n \|h_i^{(1)} - h_i^{(2)}\|^2 \quad (7)$$

Using this formulation, we define the total loss for the editing procedure as:

$$\mathcal{L}_{\text{total}} = \underbrace{\text{MSE}(h_{\ell^*}^{\text{pert}}, h_{\ell^*}^{\text{ori}})}_{\text{alignment loss}} + \lambda \cdot \underbrace{\text{MSE}(h_{\ell^*}^{\text{ori-new}}, h_{\ell^*}^{\text{ori}})}_{\text{preservation loss}} \quad (8)$$

The first term enforces representational alignment between the perturbed and original prompts. In contrast, the second term regulates the parameter update by penalizing deviations from the original hidden states of P_{ori} . The hyperparameter λ governs the balance between improving robustness and maintaining the behavior of the original model. Subsequently, we used $\mathcal{L}_{\text{total}}$ to edit $W_{\ell^*}^V$ through back propagation:

$$\begin{aligned} \mathcal{W}^{t+1} &= [W_1^{t+1}, \dots, W_{\ell^*}^{t+1}, \dots, W_N^{t+1}] \\ &= [W_1^t, \dots, W_{\ell^*}^t - \nabla_{W_{\ell^*}^V} \mathcal{L}_{\text{total}}, \dots, W_N^t], \end{aligned} \quad (9)$$

$[W_1^t, \dots, W_{\ell^*}^t, \dots, W_N^t]$ are parameters of all layers for G at t -th timestep. $W_{\ell^*}^t$ is the parameters within the key layer L_{ℓ^*} , and $\nabla_{W_{\ell^*}^V} \mathcal{L}_{\text{total}}$ is the gradient for $W_{\ell^*}^t$ at t -th timestep. After T gradient update steps, we obtain the final edited parameter set \mathcal{W}' , where only $W_{\ell^*}^V$ has been modified.

4 Experimental Setup

4.1 Baseline

To evaluate the effectiveness of our approach, we compare it against four baselines: two robustness enhancement methods and two knowledge editing methods.

4.1.1 Self-Denoising [3]. is a prompt-based technique designed to improve the resilience of LLMs to instruction-level perturbations. The method guides the model itself to recover the clean version of a perturbed input before execution. In our experiments, we employ the iterative variant (SDi), where the model is employed multiple times to progressively reduce noise in the input prompt. We select SDi because it outperforms the standard self-denoising approach [3]. However, it does not improve the inherent robustness of the model but instead relies on optimizing the input prompt. In real-world scenarios, we cannot assume that users will modify or refine their prompts prior to interacting with an LLM.

4.1.2 LoRA Fine-tuning [3]. is a robustness enhancement strategy based on representation alignment. It inserts LoRA modules into a frozen LLM and trains them to align the hidden representations of perturbed prompts with those of their unperturbed counterparts. Given pairs of original and perturbed prompts, the model minimizes the cosine distance between their mean-pooled middle-layer representations. The LoRA modules are activated only during perturbed inputs, enabling efficient robustness tuning without full model fine-tuning or ground-truth code supervision.

4.1.3 DINM [61]. is a knowledge editing method developed for detoxification tasks. It identifies the most toxic layer by comparing the hidden states of safe and unsafe completions for the same adversarial input, and fine-tunes this region using a single safe reference. To preserve general capability, it imposes a constraint on unrelated prompts. While both DINM and our method apply localized parameter updates, DINM depends on gold-standard responses, making it well-suited for tasks with clearly defined output references. In contrast, code generation lacks fixed targets, necessitating improvements in robustness without access to predefined ground-truth outputs.

4.1.4 ROME [45]. is a knowledge editing method that performs immediate and targeted parameter updates to factual associations in LLMs. It identifies the most influential layer and neuron for a given fact using causal tracing, and applies a rank-one update to the MLP output weights at that layer to inject new knowledge. In our work, we adapt ROME to the robustness enhancement setting by using a perturbed prompt and the corresponding standard answer to guide the edit.

4.2 Dataset

To evaluate the effectiveness of our method, we use code generation datasets along with their corresponding perturbed versions. Specifically, we use two widely adopted datasets:

- **HumanEval [8]** is a benchmark consisting of 164 hand-written Python programming problems. Each problem includes an input prompt, a reference solution, and a set of test cases. The prompt comprises a function signature, a functional description, and several example outputs.

- **MBPP [4]** consists of 974 Python coding problems, including a manually verified subset of 427 problems curated by the authors. Each problem includes a prompt containing a functional description, a reference solution, and test cases. We leverage the manually verified subset to conduct our study.

Based on a literature review and an online survey of practitioners, Chen et al.'s work [7] categorizes common prompt perturbations and propose an automated framework, **NLPerturbator**, which applies each type of perturbation to a given set of prompts. We follow their work and get the perturbed dataset (i.e., HumanEval-R and MBPP-R). As shown in Table 1, we adopt 20 perturbation types that span lexical and syntactic variations. These perturbations are designed to preserve the original semantics while introducing surface-level deviations that challenge the robustness of models.

To ensure a meaningful robustness evaluation, it is essential to identify those prompts that genuinely cause deviations in the output behavior of models. We first compute the pass rate (i.e., functional correctness measured via test case success) of the original prompt set, denoted as Acc_{ori} . For each perturbation type, we then compute Acc_{pert} on the corresponding perturbed prompts. The robustness drop for each prompt is defined as:

$$\Delta_{robustness} = Acc_{ori} - Acc_{pert} \quad (10)$$

However, due to the inherent randomness in LLM generation, small fluctuations in pass rate may not reliably indicate a robustness issue. To mitigate this effect and focus on impactful perturbation prompts, we introduce a filtering threshold $\delta = 0.3$ and retain only those samples for which $\Delta_{robustness} \geq \delta$. This threshold balances sensitivity and coverage: a lower value may admit noise due to random variation, while a higher threshold (e.g., $\delta = 0.5$) substantially reduces the number of usable samples. The filtered subset serves as a perturbation-sensitive benchmark for evaluating robustness-enhancing methods.

4.3 Evaluation Metrics

To evaluate the robustness of code LLMs under prompt perturbations, we employ two complementary metrics: pass@1 and Generalized Relative Improvement Ratio (G-RIR).

Pass@1 is a standard metric for evaluating the functional correctness of generated code [8]. It measures the proportion of generated outputs that pass all test cases on the first attempt. In this study, we adopt pass@1 rather than higher- k variants (e.g., pass@5 or pass@10), as it reflects the practical setting in which users expect a correct solution from a single generation without manual selection or reranking. Formally, let n denote the number of generated samples for a given prompt, c the number of correct generations. The unbiased pass@k metric is defined as:

$$Pass@k = 1 - \frac{\binom{n-c}{k}}{\binom{n}{k}}, \quad \text{for } k \leq n \quad (11)$$

G-RIR is a metric we propose to quantify the generalization ability of a robustness enhancement strategy. Given a set of $N+x$ prompts associated with a specific perturbation type, we use x prompts to update the model (e.g., $x = 1$ in our method). We then evaluate the resulting robustness improvement on the remaining

Table 1: Categories and Examples of Perturbations

ID	Name	Example (Original → Perturbed)
Addition (A1 - A3)		
A1	Extra Space inside Words	Write a python function to check ... → Write a python function to ch eck ...
A2	Repeated Words	Write a python function to check ... → Write a python python function to check ...
A3	Repeated Chars	Write a python function to check ... → Write a pyython function to check ...
Deletion (D1 - D4)		
D1	Char Deletion	Write a python function to check ... → Write a pthon function to check ...
D2	Preposition Deletion	Write a python function to check ... → Write a python function_check ...
D3	Determiner Deletion	... to check if the given number is ... → ... to check if _given number is ...
D4	Space Deletion	Write a python function to check ... → Write a python functionto check ...
Editing (E1 - E6)		
E1	Keyboard Typo	Write a python function to check ... → Write a python function to chack ...
E2	Extra Capital Letter	Write a python function to check ... → Write a python function to cHeck ...
E3	Grammatical Person Variation	Write a python function to check ... → Writes a python function to check ...
E4	Active/Passive Voice Variation	Write a python function to check ... → A python function is written to check ...
E5	Word Class Variation	Write a python function to check ... → Writer a python function to check ...
E6	Synonym Substitution	Write a python function to check ... → Write a python function to determine ...
Swap (S1 - S2)		
S1	Swap Adjacent Chars	Write a python function to check ... → Write a python fuction to check ...
S2	Swap Adjacent Words	Write a python function to check ... → Write a python to function check ...
Paraphrasing (P1 - P2)		
P1	Rephrasing Sentence	Print even numbers from a list of numbers. → Given a list of numbers, print the even numbers.
P2	Declarative to Interrogative	Print even numbers from a list of numbers. → Can you print even numbers from a list of numbers?
Combinations (C1 - C3)		
C1	A1 + E1	Write a python function to check ... → Write a py thon function to check ...
C2	A3 + E1	Write a python function to check ... → Write a pytthon function to chuck ...
C3	D1 + E1	Write a python function to check ... → Write a pythn function to chevk ...

N prompts, which exhibit the same type of perturbation but are unseen during the model updating. Formally, G-RIR is defined as:

$$\text{G-RIR} = \frac{1}{N} \sum_{i=1}^N \frac{A_{\text{pert},i}^{\text{after}} - A_{\text{pert},i}^{\text{before}}}{A_{\text{orig},i}^{\text{before}} - A_{\text{pert},i}^{\text{before}} + \epsilon} \quad (12)$$

$A_{\text{pert},i}^{\text{before}}$ and $A_{\text{pert},i}^{\text{after}}$ represent the pass@1 scores of the i -th perturbed prompt before and after applying the robustness enhancement method, respectively. $A_{\text{orig},i}^{\text{before}}$ denotes the pass@1 score of the corresponding original prompt. A small constant ϵ is added to the denominator to prevent division by zero. The numerator captures the absolute performance improvement on the perturbed prompt due to the enhancement method, while the denominator reflects the maximum possible improvement. This metric quantifies the average restoration effectiveness, normalized by the robustness gap prior to editing. A higher G-RIR indicates stronger and more consistent robustness improvements across tasks within a given perturbation type.

4.4 Experimental Setting

All experiments are conducted using PyTorch and HuggingFace Transformers on a machine equipped with an NVIDIA A800 GPU (80 GB memory). For evaluation, we use two widely adopted open-source code LLMs: CodeLlama-7b and Qwen2.5-Coder-7B, both loaded in half-precision (i.e., FP16) with left-padded tokenization.

During robustness editing, we set the learning rate to 1×10^{-3} and allow up to 20 editing steps. The Adam optimizer is used with

a weight decay of 1×10^{-5} . Early stopping is applied: the editing terminates if the loss does not improve for three consecutive steps. The regularization coefficient λ for preserving the original hidden state is set to 0.1. Only the MLP output projection matrix W_{ℓ}^V at the identified key layer L_{ℓ^*} is updated, while all other model parameters remain frozen.

For code generation, we set the sampling temperature to 0.2 to reduce randomness and ensure evaluation stability. We apply our editing method to one perturbed prompt per perturbation type and assess its generalization to other prompts of the same type.

To evaluate robustness generalization, we adopt a leave-one-in evaluation protocol within each perturbation type. Specifically, for a given type (e.g., C3 or D2) in Table 1, each task is used in turn to perform model editing or enhancement, after which the updated model is evaluated on the remaining tasks in the same category. We report performance using functional correctness (Pass@1) and generalization ability (G-RIR) to quantify robustness improvement.

5 Results and Analysis

In this paper, we aim to answer the following four research questions:

RQ.1 (Effectiveness) How effective is our method compared to existing robustness enhancement and knowledge editing approaches?

RQ.2 (Generality) How does the effectiveness of our method vary across different types of prompt perturbations?

RQ.3 (Ablation Study) How do the analyses described in Section 3 contribute to the overall effectiveness?

Table 2: Comparison of Pass@1 across models, datasets, and perturbation types. Raw scores are followed by percentage improvements over the perturbed baseline. Higher is better.

Method	CodeLlama + HumanEval		CodeLlama + MBPP		QWenCoder + HumanEval		QWenCoder + MBPP		Avg↑
	C3↑	D2↑	C3↑	D2↑	C3↑	D2↑	C3	D2↑	
Base LLM (non-perturbation)	0.600	0.583	0.688	0.710	0.739	0.625	0.780	0.769	-
Base LLM (perturbation)	0.220	0.150	0.208	0.190	0.300	0.238	0.190	0.375	-
Self-Denoising	0.062 (-72%)	0.113 (-25%)	0.033 (-84%)	0.060 (-68%)	0.014 (-95%)	0.180 (-24%)	0.444 (134%)	0.505 (35%)	-25%
LoRa	0.294 (34%)	0.283 (89%)	0.329 (58%)	0.327 (72%)	0.379 (26%)	0.198 (-17%)	0.427 (125%)	0.477 (27%)	52%
ROME	0.196 (-11%)	0.280 (87%)	0.129 (-38%)	0.201 (6%)	0.400 (33%)	0.307 (29%)	0.299 (57%)	0.546 (46%)	26%
DINM	0.227 (3%)	0.223 (49%)	0.283 (36%)	0.256 (35%)	0.177 (-41%)	0.166 (-30%)	0.392 (106%)	0.408 (9%)	21%
CREME (Ours)	0.304 (38%)	0.287 (91%)	0.340 (63%)	0.331 (74%)	0.422 (41%)	0.354 (49%)	0.396 (108%)	0.548 (46%)	64%

Table 3: Comparison of G-RIR across models, datasets, and perturbation types. Higher is better.

Method	CodeLlama + HumanEval		CodeLlama + MBPP		QWenCoder + HumanEval		QWenCoder + MBPP		Avg↑
	C3↑	D2↑	C3↑	D2↑	C3↑	D2↑	C3	D2↑	
Self-Denoising	-	-	-	-	-	-	-	-	-
LoRa	0.2407	0.3139	0.3117	0.3776	0.1218	-0.0818	0.4660	0.3445	0.2618
ROME	0.0874	0.2833	-0.2070	0.0063	0.2693	0.1934	0.2375	0.4891	0.1699
DINM	-0.0072	0.1500	0.1685	0.1207	-0.2829	-0.2083	0.3448	0.1236	0.0510
CREME (Ours)	0.2956	0.3333	0.3414	0.3786	0.3502	0.3051	0.4332	0.5120	0.3687

RQ.4 (Layer Distribution Analysis) How are robustness-sensitive layers distributed across models and perturbation types?

5.1 RQ1: Effectiveness

To evaluate the effectiveness of CREME, we compare our method with several baselines, including two robustness enhancement techniques (i.e., Self-Denoising and LoRA fine-tuning) and two knowledge editing methods (i.e., ROME and DINM). We assess performance on two representative perturbation types, C3 and D2, which represent different levels of perturbation intensity: D2 introduces a single localized modification, whereas C3 involves multiple dispersed changes. We report both absolute functional accuracy (Pass@1) and robustness generalization (G-RIR) under the leave-one-in evaluation protocol (see Section 4.4).

👍 **Pass@1 Comparison:** CREME exhibits strong potential in enhancing model robustness. As shown in Table 2, it achieves the highest Pass@1 scores in 7 out of 8 settings. In all cases, CREME outperforms the unedited model on perturbed prompts, improving the ability of models to generate functionally correct code despite input perturbations. Notably, the Self-Denoising approach performs poorly in most settings, except for QWenCoder+MBPP, contrasting with its previously reported effectiveness [3]. This discrepancy may be attributed to two factors: ❶ unlike like QwenCoder, CodeLlama is not instruction-tuned, which limits its ability to revise perturbed prompts based on meta-instructions; ❷ prompts in HumanEval average 231.6 characters in length, significantly longer than those in MBPP (88.2 characters), which increases the likelihood of semantic drift during denoising.

👍 **G-RIR Comparison:** CREME exhibits strong generalization capabilities for robustness enhancement. As shown in Table 3, it achieves the highest G-RIR scores in nearly all evaluated settings. For instance, CREME achieves a G-RIR of 0.5120 under the Qwen + MBPP + D2 configuration, significantly outperforming all baselines. Self-Denoising is excluded from the G-RIR comparison as it is designed to enhance robustness on a single instance and does not support generalization across tasks within the same perturbation

category. Additionally, some methods (e.g., ROME and DINM) yield negative G-RIR values in certain settings, suggesting that editing with a single standard answer may lead to overfitting and reduced robustness on the same perturbation.



Answer to RQ1: CREME achieves the best overall performance, with a 63% average improvement in Pass@1 over the base LLM on perturbed prompts and a 23% relative gain over the best baseline (LoRA). It also obtains the highest average G-RIR score (0.37), representing a 41% improvement over the best-performing baseline. This suggests that robustness improvements from a single edit generalize well to other tasks within the same perturbation category.

5.2 RQ2: Generality

To evaluate the generalization capability of our method across different categories of prompt perturbations, we calculate the G-RIR for each perturbation type listed in Table 1. Table 4 reports the G-RIR scores for 20 perturbation types, including additions (A1–A3), deletions (D1–D4), edits (E1–E6), swaps (S1–S2), paraphrases (P1–P2), and co-occurring categories (C1–C3). Our key findings are summarized as follows:

🔍 **Finding 1.** Our method achieves an average G-RIR score of 30% and consistently yields positive G-RIR scores across all perturbation types, indicating strong generalization.

🔍 **Finding 2.** CREME demonstrates stronger generalization on word-level perturbations compared to sentence-level rephrasings. Specifically, perturbation types in the Addition (A1–A3), Deletion (D1–D4), and Editing (E1–E6) categories generally yield higher G-RIR scores, with many cases exceeding 30%. These perturbations introduce changes that preserve the original semantic intent. In contrast, rephrasing (P1, P2) and swapping (S1, S2) perturbations are more challenging. For example, P1 yields the lowest average G-RIR scores of 16%. These types often involve sentence-level rewording or syntactic restructuring, which may cause the model to shift its

Table 4: G-RIR scores (%) across different types of prompt perturbations. Higher is better.

Perturbation Type	A1	A2	A3	D1	D2	D3	D4	E1	E2	E3	E4	E5	E6	S1	S2	P1	P2	C1	C2	C3	Avg
CodeLlama+humaneval	22	17	11	37	33	50	18	30	29	11	3	23	39	23	6	7	19	46	16	30	24
CodeLlama+MBPP	20	72	40	14	38	34	52	45	50	40	35	53	44	56	24	27	35	19	37	34	38
Qwen+humaneval	30	14	34	31	31	17	17	16	42	50	34	22	23	13	26	10	9	11	9	35	24
Qwen+MBPP	43	67	52	17	51	29	49	38	22	35	21	39	27	29	38	19	34	10	36	43	35
Average	29	43	34	25	38	33	34	33	36	34	23	34	33	30	24	16	24	22	25	36	30

Table 5: Cross-Category Generalization: G-RIR (%) on Unseen Perturbation Types After Editing on C3 (HumanEval+QwenCoder)

Perturbation Type	A1	A2	A3	D1	D2	D3	D4	E1	E2	E3	E4	E5	E6	S1	S2	P1	P2	C1	C2	Avg
G-RIR	24	19	28	23	21	2	16	9	32	38	25	19	20	3	19	10	15	11	7	23

internal attention or misinterpret the functional objective of the prompt.

🔍 **Finding 3.** The effectiveness of our method varies across datasets. The average G-RIR score on MBPP is consistently higher than that on HumanEval for both CodeLlama and Qwen. For example, CodeLlama achieves an average G-RIR of 38% on MBPP compared to 24% on HumanEval, while Qwen scores 35% on MBPP and 24% on HumanEval. This discrepancy may be attributed to the nature of the tasks: MBPP prompts are generally shorter and more templated, whereas HumanEval problems tend to be more descriptive and structurally diverse. Consequently, CREME demonstrates better generalization on simpler, more regular prompts.

🔍 **Finding 4.** To evaluate whether model editing tailored to a specific perturbation category can generalize to others, we conduct an additional cross-category evaluation. Specifically, the model is edited using only C3 type examples under the HumanEval+QwenCoder setting and then tests across all other perturbation types. As shown in Table 5, the model achieves an average G-RIR improvement of 18% on unseen perturbation types, compared to 24% when both editing and evaluation are performed within the same category (as reported in Table 4). These results suggest that CREME does not overfit to a specific perturbation category and can generalize to a broader range of prompt variations.

💡 **Answer to RQ2:** Our method demonstrates strong generalization across various categories of perturbations, achieving an average improvement of 30%. It performs better on simpler perturbations (e.g., word-level modifications) than on more complex sentence-level rephrasings. Moreover, CREME preserves robustness improvements even when trained on a single perturbation type and evaluated across others, achieving an average cross-type G-RIR of 18%.

5.3 RQ3: Ablation Study

To investigate the contribution of each component to the effectiveness of CREME, we create the following variants:

- **CREME_Full:** The complete version of our method, exactly as Section 3 illustrated.
- **w/o Layer Localization:** This variant skips the causal tracing step and instead randomly selects a middle layer for editing. This helps assess the importance of precisely locating the key layer responsible for robustness degradation.

- **w/o Early Stopping:** Early stopping is disabled, allowing all editing steps to be executed regardless of the loss trend. This removes the safeguard designed to prevent overfitting.
- **w/o Preserve Loss:** This variant removes the preservation loss term, which constrains the model to maintain its behavior on clean inputs, thereby allowing the model to deviate from its original performance.

Table 6 reports the G-RIR scores of CREME and its ablated variants across four evaluation settings. Overall, CREME_Full consistently achieves the highest G-RIR scores, validating the contribution of each individual component. Among the ablation variants, removing the layer localization module results in the greatest performance degradation (20.4% on average), underscoring the critical importance of locating the key layer related to robustness. Disabling early stopping results in a 14.2% average decline, emphasizing its importance in preventing overfitting during model editing. Removing the preservation loss term causes a 12.8% reduction, indicating that preserving model behavior on clean inputs is also crucial for achieving robust generalization.

💡 **Answer to RQ3:** All components of CREME are essential to its effectiveness, with layer localization contributing most significantly (−20.4% G-RIR).

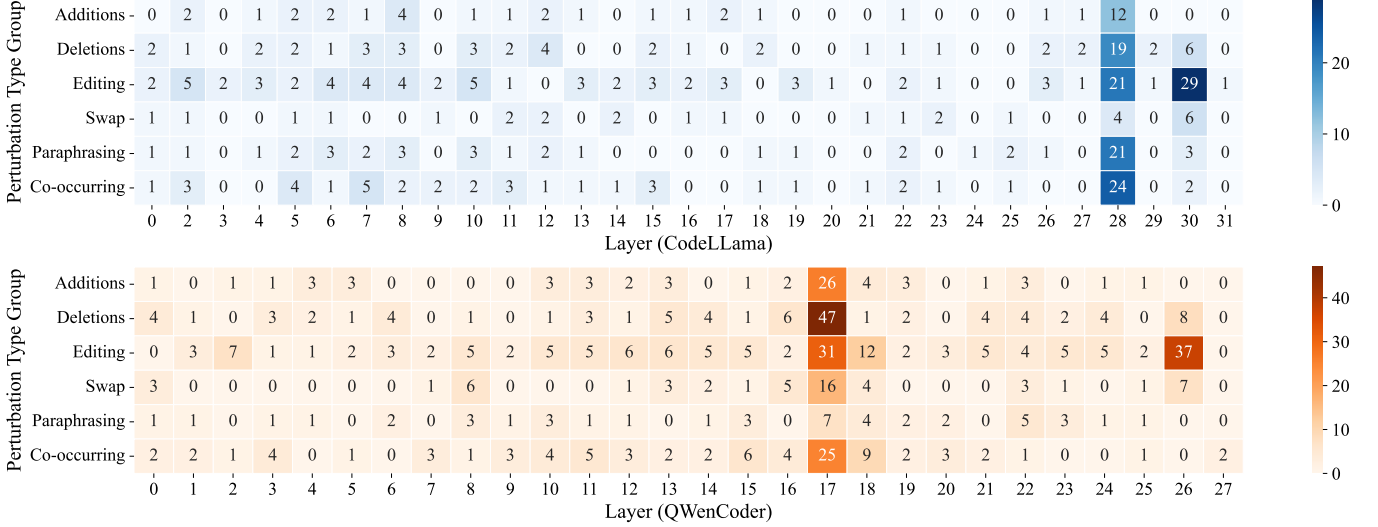
5.4 RQ4: Layer Distribution Analysis

To understand how robustness-related information is localized across the model’s architecture, we analyze the distribution of key layers selected by our causal tracing procedure under different perturbation types and models. Figure 3 presents the heatmaps of key layer frequencies for the CodeLlama-7b and Qwen2.5-Coder-7B models, grouped by six perturbation categories. The “key layer frequency” refers to how many times each transformer layer is selected as the most robustness-sensitive layer across all editing cases (2 datasets × 20 perturbation types).

🔗 **Model Dimension:** Robustness-sensitive layers are concentrated in specific regions of the network and vary significantly across models. In the case of CodeLlama, key layers are generally located in the later stages of the model and exhibit a more dispersed distribution (mean = 18.93, standard deviation = 10.08). Notably, layers 28 and 30 are the most frequently identified, together accounting for over 40% of all key layers. In contrast, QwenCoder exhibits a more centralized distribution (mean = 15.76, standard deviation = 6.57), with a strong concentration at layer 17, suggesting a more

Table 6: G-RIR scores for the ablation variants of our method under different settings. Higher is better.

Method	CodeLlama + HumanEval		CodeLlama + MBPP		QWenCoder + HumanEval		QWenCoder + MBPP	
	C3	D2	C3	D2	C3	D2	C3	D2
w/o Layer Localization	0.2296↓	0.2306	0.3003↓	0.2364↓	0.3249↓	0.2594↓	0.3074↓	0.4838↓
w/o Early Stopping	0.2593	0.2278↓	0.3305	0.2824	0.3540	0.2634	0.3398	0.4969
w/o Preserve Loss ($\mathcal{L}_{\text{preserve}}$)	0.2481	0.2528	0.3469	0.3042	0.3383	0.2693	0.3467	0.4870
CREME_Full (Ours)	0.2768	0.3194	0.3414	0.3786	0.3502	0.3275	0.4766	0.5120

**Figure 3: Key Layer Distribution by Perturbation Type Group**

localized region for robustness-relevant computations. Despite differences in the number of layers and architectural design, both models consistently exhibit a concentration of robustness-sensitive layers in the middle-to-deep regions of their networks. These findings support the hypothesis that the robustness-related region is not uniformly distributed across the model but is instead governed by specific layers that encode perturbation-sensitive representations. Identifying and targeting these regions is therefore essential for effective robustness enhancement.

✦ Perturbation Dimension: Most perturbation types exhibit similar distributions of robustness-sensitive key layers, while certain categories display distinct patterns. In the case of CodeLlama, most perturbation types (e.g., additions, deletions, and paraphrasing) tend to concentrate key layers around Layer 28. However, the Editing category exhibits a distinct pattern by shifting the key layer distribution toward Layer 30. This shift suggests that editing perturbations may engage robustness-sensitive mechanisms in deeper layers, potentially due to their grammatical and lexical complexity that demands more semantic-level processing. A similar trend is observed in QWenCoder, where most perturbation types concentrate around Layer 17. However, Editing perturbations notably shift the distribution toward Layer 26. These results suggest that while key layers remain relatively consistent within similar perturbation types, distinct perturbation categories can activate robustness-sensitive regions in different parts of the model.

Overall, these findings suggest that robustness-sensitive regions are not uniformly distributed across all layers but tend to cluster within specific parts of the network. The locations of these regions

vary depending on both the model architecture and the type of perturbation. This non-uniformity implies the existence of specialized subregions within the model that are particularly responsible for processing robustness-related information. These observations empirically validate our layer localization strategy and offer valuable insights for designing more effective robustness enhancement techniques.



Answer to RQ4: Robustness-sensitive layers are unevenly distributed and tend to cluster within specific layers of the model, with their locations varying across model architectures and perturbation types.

6 Discussion

6.1 Impact of Editing on Model Performance

While CREME effectively enhances robustness against perturbed prompts, it is crucial to evaluate whether the editing process inadvertently degrades model performance on clean data. We conduct a case study by randomly sampling 40 editing tasks from the dataset used in each experimental configuration. To ensure balanced coverage across perturbation types, the sampled tasks include two instances for each perturbation type listed in Table 1. For each task, we apply the CREME editing procedure and then evaluate the edited model on the corresponding original dataset (i.e., HumanEval or MBPP) with clean prompts. Table 7 presents the comparison of pass@1 and pass@10 accuracy before and after editing. The results indicate that the performance of CREME remains stable on clean inputs, with most variations within a $\pm 1\%$ margin.

Table 7: Code accuracy (*pass@k*) on original inputs before and after editing. Δ denotes the change in accuracy after editing.

Model	Dataset	pass@1 (Orig)	pass@1 (Edited)	Δ	pass@10 (Orig)	pass@10 (Edited)	Δ
CodeLlama-7b	HumanEval	30.43%	30.25%	-0.17%	43.90%	44.62%	0.72%
CodeLlama-7b	MBPP	51.69%	51.67%	-0.02%	67.33%	67.35%	0.02%
Qwen2.5-Coder-7B	HumanEval	45.67%	46.35%	0.68%	62.80%	62.56%	-0.24%
Qwen2.5-Coder-7B	MBPP	63.81%	63.59%	-0.22%	79.10%	79.58%	0.48%

Overall, these results demonstrate that CREME achieves robustness enhancement with minimal impact on the original capabilities of the model.

6.2 Costs of CREME

To assess the practical feasibility of CREME, we conducted a cost evaluation on the HumanEval+QwenCoder setting using the C3 perturbation category. Table 8 presents a comparison of CREME with four representative baseline methods in terms of localization, training, and testing time.

The layer localization step of CREME takes an average of 495s (8.3 minutes). While this step introduces additional overhead, it is crucial for identifying the most robustness-sensitive layer. In contrast, the two model editing methods, ROME and DINM, rely on either fixed layers or hidden-state distance, neither of which is robustness-aware. As demonstrated in Table 6, our localization strategy substantially improves robustness over fixed-layer editing, justifying the additional cost. Regarding training time, CREME requires only 5.3s, which is significantly less than DINM (15.8s) and comparable to the others. For inference, CREME’s average test time (3.0s) is competitive and markedly faster than Self-Denoising (29.4s), which iteratively adjusts prompts during generation. Given that models are typically deployed across many downstream tasks after editing, minimizing test time is more critical.

To reduce the overhead of layer localization, we propose a lightweight variant that limits localization to a small set of frequently identified robustness-sensitive layers from Section 5.4 (e.g., layers 28/30 in CodeLlama and 17/26 in QwenCoder), which significantly lowers localization time cost while maintaining effectiveness.

Table 8: Comparison of Time Costs (s) for CREME and Baselines on HumanEval+QwenCoder (C3 perturbation)

Method	Localization Time (s)	Train Time (s)	Test Time (s)
Self-Denoising	-	-	29.4
LoRa	-	9.2	2.8
ROME	-	7.59	3.4
DINM	0.2	15.8	3.9
CREME	494.5	5.3	3.0

6.3 Hyperparameter Sensitivity of λ

The hyperparameter λ in the Equation 8 controls the trade-off between the preservation loss and the alignment loss. To guide practitioners in choosing an appropriate value for λ , we conducted a sensitivity analysis using the C3 perturbation category under the HumanEval+QwenCoder setting. As shown in Table 9, both G-RIR and Pass@1 scores improve as λ increases from 0.01 to 0.1, peaking

at $\lambda = 0.1$. Beyond this point, performance begins to decline. These findings indicate that $\lambda = 0.1$ provides a robust trade-off and can serve as a strong default for future applications of our framework.

Table 9: Sensitivity Analysis of λ on HumanEval+QwenCoder (C3 perturbation)

$\lambda(\text{lambda})$	0.01	0.05	0.1	0.2	0.5	1.0
G-RIR (%)	32.1	33.2	35.0	30.3	32.8	27.6
Pass@1 (%)	41.4	41.9	42.2	40.8	41.2	39.6

6.4 Threats to Validity

6.4.1 Internal Validity. A potential threat arises from the accuracy of our key layer localization strategy based on causal tracing. Although we identify robustness-sensitive layers using both restoration improvement metric and L2-based refinement, there remains a risk of misidentification due to randomness in model behavior or dataset noise. To mitigate this, we aggregate results across multiple tasks and perturbation types. Another concern is the stability of evaluation. Since code generation involves stochastic sampling (i.e., with temperature = 0.2), the results may exhibit slight variance. We address this by standardizing generation settings and averaging results across multiple completions per prompt.

6.4.2 External Validity. Our study is conducted using two open-source code generation models (i.e., CodeLlama and QwenCoder) and two widely used benchmarks (i.e., HumanEval and MBPP). While these choices reflect diversity in both model architectures and prompt styles, the generalizability of our approach to other LLMs (e.g., GPT-4 and DeepSeek) or to domains beyond code generation (e.g., natural language question answering or summarization) remains unverified. Our concern is that commercial LLMs are continuously updated and may become outdated or permanently inaccessible. In contrast, open-source LLMs offer stable access once released, and their historical versions can be revisited. We select code LLMs for our study because they can generate code directly without requiring additional fine-tuning on external code datasets. Furthermore, the perturbations employed are derived from the NLPerturbator taxonomy, which may not comprehensively represent robustness challenges under more adversarial or out-of-distribution conditions. Extending our evaluation to include broader and more aggressive perturbation strategies will be an important direction for future work.

7 Related Work

In this section, we present a comprehensive review of prior research on LLM robustness and the knowledge editing methods. Beyond LLM robustness, SE work has studied software security and practical deployment issues [10, 31, 40, 51, 66].

7.1 LLM Robustness

Despite achieving impressive performance in increasingly sophisticated tasks [15, 16, 41], LLMs remain sensitive to input perturbations. While humans are generally robust to minor variations in natural language task descriptions [57], LLMs often produce significantly different outputs in response to such changes [23, 48], highlighting their limited robustness. Wang et al. [59] and Zhu et al. [71] introduce benchmark suites such as Adversarial GLUE and PromptRobust, which systematically evaluate LLMs under adversarial or subtly perturbed prompts.

In the domain of code generation, a series of studies have revealed that current code LLMs are sensitive to even minor variations [27, 53, 62]. Chen et al. [7] categorize 18 types of natural language perturbations along with three co-occurring combinations and develop NLPerturbator, a framework targeting real-world prompt perturbations. Their findings show that such perturbations can substantially degrade code generation performance (e.g., up to 21.2%, and 4.8% to 6.1% on average). Lin et al. [38] focus on non-functional robustness and introduce RobuNFR, a benchmark designed to evaluate LLMs under real-world noisy contexts in question-answering (QA) tasks. Their results demonstrate that LLMs are highly sensitive to such noise, thereby extending robustness analysis beyond prompt-level perturbations. Mastropaolo et al. [44] conduct an empirical study on GitHub Copilot, showing that 46% of Copilot’s outputs changed when given semantically equivalent paraphrases. This indicates a concerning lack of robustness in popular commercial models.

To improve robustness, researchers have proposed a variety of mitigation strategies. Agrawal et al. [3] introduce selective instruction augmentation and inference techniques, including self-denoising and representation alignment. They find that self-denoising achieves substantially higher performance gains than alternative strategies. Hu et al. [24] propose two mitigation strategies for retrieval-augmented generation: robust prompt alignment, which maps perturbed prompts to semantically equivalent canonical forms to stabilize retrieval, and retrieval consistency filtering, which filters out prompts yielding inconsistent retrieval results across paraphrases. Wang et al. [58] explore instruction-level noise in instruction-tuned models and find that adversarial prompts can severely degrade task performance. They further propose robustness training using paraphrased instruction variants to improve model resilience. Different from prior studies, we apply knowledge editing to enhance model robustness, without relying on additional ensemble strategies. Our method requires only a single example for intervention, offering a novel approach to improving the robustness of LLMs.

7.2 Knowledge Editing

Knowledge editing has emerged as a promising direction for updating factual knowledge in LLMs without full retraining [5, 19, 21, 25, 42, 64, 69], which can be categorized into three main paradigms [63]. External memorization-based methods [13, 20, 26, 30, 33, 34, 43, 47, 49, 68] leverage an external memory to store new knowledge without modifying the pre-trained weights, thereby fully preserving the original knowledge encoded in the LLM. Global optimization-based methods [9, 12, 18, 22, 29, 50, 54, 70] aim to integrate new knowledge into pre-trained LLMs in a generalizable manner through optimization guided by the new information. These methods employ

tailored strategies to constrain the impact on existing knowledge, distinguishing them from naive fine-tuning. The third category, local modification-based methods [11, 35, 65], is most relevant to our work. These methods aim to identify the parameters associated with specific knowledge in LLMs and selectively update them to incorporate new information related to the edit.

Meng et al. [45] propose ROME, a pioneering local editing method that identifies and updates key neuron activations within specific transformer layers responsible for storing factual associations. By analytically tracing causal dependencies, ROME directly modifies the internal representations of a model to reflect new factual information with high locality and minimal side effects. Building on this, Meng et al. [46] introduce MEMIT, a scalable framework that enables batch editing of multiple facts by optimizing a shared intervention across selected layers. Wang et al. [61] extend this line of work to the domain of safety, applying knowledge editing to mitigate toxic behaviors in LLMs by targeting offensive content and aligning representations with non-toxic alternatives. These methods underscore the potential of local interventions to precisely and efficiently influence LLM behavior. However, prior approaches have primarily focused on factual updates, whereas our work explores the novel application of local knowledge editing to improve robustness against prompt perturbations.

8 Conclusion

In this paper, we presented **CREME**, a lightweight model editing framework to enhance the robustness of LLMs against natural language prompt perturbations. By leveraging a pair of original and perturbed prompts, CREME identifies robustness-sensitive layers through causal tracing and applies targeted parameter updates to align their internal representations. We conduct extensive experiments on the HumanEval and MBPP benchmarks using two representative code generation models. The results show that CREME substantially improves functional correctness on perturbed prompts, achieving a 63% increase in Pass@1 accuracy, while maintaining performance on clean inputs within a $\pm 1\%$ margin. We further assess generalization and show that CREME restores up to 30% of robustness across perturbation types. Ablation studies confirmed the importance of each component, particularly layer localization. Our analysis of key layer distributions reveals that robustness-sensitive regions are concentrated in the middle-to-deep layers and vary with model architecture, offering actionable insights for future robustness-oriented interventions. CREME provides a practical path toward making LLMs more reliable under real-world prompt variability, without requiring full retraining or architectural modifications. Future work will explore extending CREME to broader task domains and evaluating its effectiveness against more challenging perturbations.

Acknowledgement

This research is supported by the National Key R&D Program of China (No. 2024YFB4506400) and the National Research Foundation, under its Investigatorship Grant (NRF-NRFI08-2022-0002). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore.

References

- [1] 2025. Replication Package. <https://github.com/MapleQwQ/CREME>.
- [2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).
- [3] Aryan Agrawal, Lisa Alazraki, Shahin Honarvar, and Marek Rei. 2025. Enhancing LLM robustness to perturbed instructions: An empirical study. *arXiv preprint arXiv:2504.02733* (2025).
- [4] Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, et al. 2021. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732* (2021).
- [5] Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. 2023. Leace: Perfect linear concept erasure in closed form. *Advances in Neural Information Processing Systems* 36 (2023), 66044–66063.
- [6] Junkai Chen, Xing Hu, Zhenhao Li, Cuiyun Gao, Xin Xia, and David Lo. 2024. Code search is all you need? improving code suggestions with code search. In *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering*, 1–13.
- [7] Junkai Chen, Zhenhao Li, Xing Hu, and Xin Xia. 2024. Nlperturbator: Studying the robustness of code llms to natural language variations. *arXiv preprint arXiv:2406.19783* (2024).
- [8] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374* (2021).
- [9] Sanyuan Chen, Yutai Hou, Yiming Cui, Wanxiang Che, Ting Liu, and Xiangzhan Yu. 2020. Recall and learn: Fine-tuning deep pretrained language models with less forgetting. *arXiv preprint arXiv:2004.12651* (2020).
- [10] Zirui Chen, Zhipeng Xue, Jiayuan Zhou, Xing Hu, Xin Xia, and Xiaohu Yang. 2025. Diffploit: Facilitating Cross-Version Exploit Migration for Open Source Library Vulnerabilities. *arXiv:2511.12950 [cs.SE]* <https://arxiv.org/abs/2511.12950>
- [11] Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2021. Knowledge neurons in pretrained transformers. *arXiv preprint arXiv:2104.08696* (2021).
- [12] Nicola De Cao, Wilker Aziz, and Ivan Titov. 2021. Editing factual knowledge in language models. *arXiv preprint arXiv:2104.08164* (2021).
- [13] Qingxiu Dong, Damai Dai, Yifan Song, Jingjing Xu, Zhifang Sui, and Lei Li. 2022. Calibrating factual knowledge in pretrained language models. *arXiv preprint arXiv:2210.03329* (2022).
- [14] Xueying Du, Mingwei Liu, Kaixin Wang, Hanlin Wang, Junwei Liu, Yixuan Chen, Jiayi Feng, Chaofeng Sha, Xin Peng, and Yiling Lou. 2023. Classeval: A manually-crafted benchmark for evaluating llms on class-level code generation. *arXiv preprint arXiv:2308.01861* (2023).
- [15] Sidong Feng and Chunyang Chen. 2024. Prompting is all you need: Automated android bug replay with large language models. In *Proceedings of the 46th IEEE/ACM International Conference on Software Engineering*, 1–13.
- [16] Mingyang Geng, Shangwen Wang, Dezun Dong, Haotian Wang, Ge Li, Zhi Jin, Xiaoguang Mao, and Xiangke Liao. 2024. Large language models are few-shot summarizers: Multi-intent comment generation via in-context learning. In *Proceedings of the 46th IEEE/ACM International Conference on Software Engineering*, 1–13.
- [17] Mor Geva, Avi Caciularu, Kevin Ro Wang, and Yoav Goldberg. 2022. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. *arXiv preprint arXiv:2203.14680* (2022).
- [18] Jia-Chen Gu, Hao-Xiang Xu, Jun-Yu Ma, Pan Lu, Zhen-Hua Ling, Kai-Wei Chang, and Nanyun Peng. 2024. Model editing harms general abilities of large language models: Regularization to the rescue. *arXiv preprint arXiv:2401.04700* (2024).
- [19] Akshat Gupta, Anurag Rao, and Gopala Anumanchipalli. 2024. Model editing at scale leads to gradual and catastrophic forgetting. *arXiv preprint arXiv:2401.07453* (2024).
- [20] Tom Hartvigsen, Swami Sankaranarayanan, Hamid Palangi, Yoon Kim, and Marzyeh Ghassemi. 2023. Aging with grace: Lifelong model editing with discrete key-value adaptors. *Advances in Neural Information Processing Systems* 36 (2023), 47934–47959.
- [21] Peter Hase, Mohit Bansal, Been Kim, and Asma Ghandeharioun. 2023. Does localization inform editing? surprising differences in causality-based localization vs. knowledge editing in language models. *Advances in Neural Information Processing Systems* 36 (2023), 17643–17668.
- [22] Peter Hase, Mona Diab, Asli Celikyilmaz, Xian Li, Zornitsa Kozareva, Veselin Stoyanov, Mohit Bansal, and Srinivasan Iyer. 2023. Methods for measuring, updating, and visualizing factual beliefs in language models. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 2714–2731.
- [23] Shahin Honarvar, Mark van der Wilk, and Alastair F Donaldson. 2025. Turbulence: Systematically and automatically testing instruction-tuned large language models for code. In *2025 IEEE Conference on Software Testing, Verification and Validation (ICST)*. IEEE, 80–91.
- [24] Zhibo Hu, Chen Wang, Yanfeng Shu, Hye-Young Paik, and Liming Zhu. 2024. Prompt perturbation in retrieval-augmented generation based large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 1119–1130.
- [25] Wenyue Hua, Jiang Guo, Mingwen Dong, Henghui Zhu, Patrick Ng, and Zhiguo Wang. 2024. Propagation and pitfalls: Reasoning-based assessment of knowledge editing through counterfactual tasks. *arXiv preprint arXiv:2401.17585* (2024).
- [26] Zeyu Huang, Yikang Shen, Xiaofeng Zhang, Jie Zhou, Wenge Rong, and Zhang Xiong. 2023. Transformer-patcher: One mistake worth one neuron. *arXiv preprint arXiv:2301.09785* (2023).
- [27] Man Ho Lam, Chaosheng Wang, Jen-tse Huang, and Michael R Lyu. 2025. CODE-CRASH: Stress Testing LLM Reasoning under Structural and Semantic Perturbations. *arXiv preprint arXiv:2504.14119* (2025).
- [28] Andrew Lee, Xiaoyan Bai, Itamar Pres, Martin Wattenberg, Jonathan K Kummerfeld, and Rada Mihalcea. 2024. A mechanistic understanding of alignment algorithms: A case study on dpo and toxicity. *arXiv preprint arXiv:2401.01967* (2024).
- [29] Kyungjae Lee, Wookje Han, Seung-won Hwang, Hwaran Lee, Joonsuk Park, and Sang-Woo Lee. 2022. Plug-and-play adaptation for continuously-updated qa. *arXiv preprint arXiv:2204.12785* (2022).
- [30] Daliang Li, Ankit Singh Rawat, Manzil Zaheer, Xin Wang, Michal Lukasik, Andreas Veit, Felix Yu, and Sanjiv Kumar. 2022. Large language models with controllable working memory. *arXiv preprint arXiv:2211.05110* (2022).
- [31] Jia Li, Zhuo Li, Huangzhao Zhang, Ge Li, Zhi Jin, Xing Hu, and Xin Xia. 2022. Poison attack and defense on deep source code processing models. *arXiv preprint arXiv:2210.17029* (2022).
- [32] Raymond Li, Loubna Ben Allal, Yangtian Zi, Niklas Muennighoff, Denis Kocetkov, Chenghao Mou, Marc Marone, Christopher Akiki, Jia Li, Jenny Chim, et al. 2023. Starcoder: may the source be with you! *arXiv preprint arXiv:2305.06161* (2023).
- [33] Shuaiyi Li, Yang Deng, Deng Cai, Hongyuan Lu, Liang Chen, and Wai Lam. 2024. Consecutive model editing with batch alongside hook layers. *arXiv e-prints* (2024), arXiv–2403.
- [34] Xiaopeng Li, Shasha Li, Bin Ji, Shezheng Song, Xi Wang, Jun Ma, Jie Yu, Xiaodong Liu, Jing Wang, and Weimin Zhang. 2024. SWEA: changing factual knowledge in large language models via subject word embedding altering. *arXiv e-prints* (2024), arXiv–2401.
- [35] Xiaopeng Li, Shasha Li, Shezheng Song, Jing Yang, Jun Ma, and Jie Yu. 2024. Pmet: Precise model editing in a transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38, 18564–18572.
- [36] Xiaopeng Li, Shangwen Wang, Shasha Li, Jun Ma, Jie Yu, Xiaodong Liu, Jing Wang, Bin Ji, and Weimin Zhang. 2024. Model Editing for LLMs4Code: How Far are We? *arXiv preprint arXiv:2411.06638* (2024).
- [37] Feng Lin, Dong Jae Kim, et al. 2024. When llm-based code generation meets the software development process. *arXiv e-prints* (2024), arXiv–2403.
- [38] Feng Lin, Dong Jae Kim, Zhenhao Li, Jinqiu Yang, et al. 2025. Robunfr: Evaluating the robustness of large language models on non-functional requirements aware code generation. *arXiv preprint arXiv:2503.22851* (2025).
- [39] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437* (2024).
- [40] Shuhan Liu, Jiayuan Zhou, Xing Hu, Filipe Roseiro Cogo, Xin Xia, and Xiaohu Yang. 2025. An empirical study on vulnerability disclosure management of open source software systems. *ACM Transactions on Software Engineering and Methodology* (2025).
- [41] Zhe Liu, Chunyang Chen, Junjie Wang, Xing Che, Yuekai Huang, Jun Hu, and Qing Wang. 2023. Fill in the blank: Context-aware automated text input generation for mobile gui testing. In *2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE)*. IEEE, 1355–1367.
- [42] Michelle Lo, Shay B Cohen, and Fazl Barez. 2024. Large language models relearn removed concepts. *arXiv preprint arXiv:2401.01814* (2024).
- [43] Aman Madaan, Niket Tandon, Peter Clark, and Yiming Yang. 2022. Memory-assisted prompt editing to improve GPT-3 after deployment. *arXiv preprint arXiv:2201.06009* (2022).
- [44] Antonio Mastropaolo, Luca Pascarella, Emanuela Guglielmi, Matteo Ciniselli, Simone Scalabrino, Rocco Oliveto, and Gabriele Bavota. 2023. On the robustness of code generation techniques: An empirical study on github copilot. In *2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE)*. IEEE, 2149–2160.
- [45] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. *Advances in neural information processing systems* 35 (2022), 17359–17372.
- [46] Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. 2022. Mass-editing memory in a transformer. *arXiv preprint arXiv:2210.07229* (2022).
- [47] Eric Mitchell, Charles Lin, Antoine Bosselut, Christopher D Manning, and Chelsea Finn. 2022. Memory-based model editing at scale. In *International Conference on Machine Learning*. PMLR, 15817–15831.

- [48] Milad Moradi and Matthias Samwald. 2021. Evaluating the robustness of neural language models to input perturbations. *arXiv preprint arXiv:2108.12237* (2021).
- [49] Shikhar Murty, Christopher D Manning, Scott Lundberg, and Marco Tulio Ribeiro. 2022. Fixing model bugs with natural language patches. *arXiv preprint arXiv:2211.03318* (2022).
- [50] Shiwen Ni, Dingwei Chen, Chengming Li, Xiping Hu, Ruifeng Xu, and Min Yang. 2023. Forgetting before learning: Utilizing parametric arithmetic for knowledge updating in large language models. *arXiv preprint arXiv:2311.08011* (2023).
- [51] Shengyi Pan, Jiayuan Zhou, Filipe Roseiro Cogo, Xin Xia, Lingfeng Bao, Xing Hu, Shanping Li, and Ahmed E Hassan. 2022. Automated unearthing of dangerous issue reports. In *Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. 834–846.
- [52] Xinyue Shen, Zeyuan Chen, Michael Backes, and Yang Zhang. 2023. In chatgpt we trust? measuring and characterizing the reliability of chatgpt. *arXiv preprint arXiv:2304.08979* (2023).
- [53] Atsushi Shirafuji, Yutaka Watanobe, Takumi Ito, Makoto Morishita, Yuki Nakamura, Yusuke Oda, and Jun Suzuki. 2023. Exploring the robustness of large language models for solving programming problems. *arXiv preprint arXiv:2306.14583* (2023).
- [54] Anton Sinitin, Vsevolod Plokhoniyuk, Dmitriy Pyrkun, Sergei Popov, and Artem Babenko. 2020. Editable neural networks. *arXiv preprint arXiv:2004.00345* (2020).
- [55] Jiuding Sun, Chantal Shaib, and Byron C Wallace. 2023. Evaluating the zero-shot robustness of instruction-tuned language models. *arXiv preprint arXiv:2306.11270* (2023).
- [56] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).
- [57] Candace Walkington, Virginia Clinton, and Anthony Sparks. 2019. The effect of language modification of mathematics story problems on problem-solving in online homework. *Instructional Science* 47, 5 (2019), 499–529.
- [58] Bin Wang, Chengwei Wei, Zhengyuan Liu, Geyu Lin, and Nancy F Chen. 2024. Resilience of large language models for noisy instructions. *arXiv preprint arXiv:2404.09754* (2024).
- [59] Boxin Wang, Chejian Xu, Shuohang Wang, Zhe Gan, Yu Cheng, Jianfeng Gao, Ahmed Hassan Awadallah, and Bo Li. 2021. Adversarial glue: A multi-task benchmark for robustness evaluation of language models. *arXiv preprint arXiv:2111.02840* (2021).
- [60] Jindong Wang, Xixu Hu, Wenxin Hou, Hao Chen, Runkai Zheng, Yidong Wang, Linyi Yang, Haojun Huang, Wei Ye, Xiubo Geng, et al. 2023. On the robustness of chatgpt: An adversarial and out-of-distribution perspective. *arXiv preprint arXiv:2302.12095* (2023).
- [61] Mengru Wang, Ningyu Zhang, Ziwen Xu, Zekun Xi, Shumin Deng, Yunzhi Yao, Qishen Zhang, Linyi Yang, Jindong Wang, and Huajun Chen. 2024. Detoxifying large language models via knowledge editing. *arXiv preprint arXiv:2403.14472* (2024).
- [62] Shiqi Wang, Zheng Li, Haifeng Qian, Chenghao Yang, Zijian Wang, Mingyue Shang, Varun Kumar, Samson Tan, Baishakhi Ray, Parminder Bhatia, et al. 2022. ReCode: Robustness evaluation of code generation models. *arXiv preprint arXiv:2212.10264* (2022).
- [63] Song Wang, Yaochen Zhu, Haochen Liu, Zaiyi Zheng, Chen Chen, and Jundong Li. 2024. Knowledge editing for large language models: A survey. *Comput. Surveys* 57, 3 (2024), 1–37.
- [64] Yifan Wei, Xiaoyan Yu, Huanhuan Ma, Fangyu Lei, Yixuan Weng, Ran Song, and Kang Liu. 2023. Assessing knowledge editing in language models via relation perspective. *arXiv preprint arXiv:2311.09053* (2023).
- [65] Xinwei Wu, Junzhuo Li, Minghui Xu, Weilong Dong, Shuangzhi Wu, Chao Bian, and Deyi Xiong. 2023. Depn: Detecting and editing privacy neurons in pretrained language models. *arXiv preprint arXiv:2310.20138* (2023).
- [66] Yanming Yang, Xing Hu, Zhipeng Gao, Jinfu Chen, Chao Ni, Xin Xia, and David Lo. 2024. Federated learning for software engineering: A case study of code clone detection and defect prediction. *IEEE Transactions on Software Engineering* 50, 2 (2024), 296–321.
- [67] Ningyu Zhang, Yunzhi Yao, Bozhong Tian, Peng Wang, Shumin Deng, Mengru Wang, Zekun Xi, Shengyu Mao, Jintian Zhang, Yuansheng Ni, et al. 2024. A comprehensive study of knowledge editing for large language models. *arXiv preprint arXiv:2401.01286* (2024).
- [68] Ce Zheng, Lei Li, Qingxiu Dong, Yuxuan Fan, Zhiyong Wu, Jingjing Xu, and Baobao Chang. 2023. Can we edit factual knowledge by in-context learning? *arXiv preprint arXiv:2305.12740* (2023).
- [69] Zexuan Zhong, Zhengxuan Wu, Christopher D Manning, Christopher Potts, and Danqi Chen. 2023. Mquake: Assessing knowledge editing in language models via multi-hop questions. *arXiv preprint arXiv:2305.14795* (2023).
- [70] Chen Zhu, Ankit Singh Rawat, Manzil Zaheer, Srinadh Bhojanapalli, Daliang Li, Felix Yu, and Sanjiv Kumar. 2020. Modifying memories in transformer models. *arXiv preprint arXiv:2012.00363* (2020).
- [71] Kaijie Zhu, Jindong Wang, Jiaheng Zhou, Zichen Wang, Hao Chen, Yidong Wang, Linyi Yang, Wei Ye, Yue Zhang, Neil Gong, et al. 2023. Promptrobust: Towards evaluating the robustness of large language models on adversarial prompts. In *Proceedings of the 1st ACM Workshop on Large AI Systems and Models with Privacy and Safety Analysis*. 57–68.