Evaluating LLMs for One-Shot Patching of Real and Artificial Vulnerabilities

Aayush Garg aayush.garg@list.lu
Zanis Ali Khan zanis-ali.khan@list.lu
Renzo Degiovanni renzo.degiovanni@list.lu
Qiang Tang qiang.tang@list.lu
Luxembourg Institute of Science and Technology (LIST), Luxembourg

Aim

To assess: RQ1. How effective are LLMs in patching real vulnerabilities vs. artificial vulnerabilities?

RQ2: How complementary and overlapping are LLMs in patching real vulnerabilities vs. artificial vulnerabilities?

Motivating example

```java
private static void decompress
 (final InputStream in, final byte[] out)
 throws IOException {
  int position = 0;
  final int total = out.length;
  while (position < total) {
   final int n = in.read();
   if (n > 128) {
    final int value = in.read();
    for (int i = 0; i < (n & 0x7f); i++) {
     out[position++] = (byte) value; }
   } else {
    for (int i = 0; i < n; i++) {
    out[position++] = (byte) in.read();}
   }
  }
 }
```

(a) Real Vulnerability (CVE-2018-17201)

```java
private static void decompress
 (final InputStream in, final byte[] out)
 throws IOException {
  int position = 0;
  final int total = out.length;
  while (position < total) {
   final int n = in.read();
   if (n < 0) {
    throw new ImageReadException("Error decompressing
          RGBE file"); }
   if (n > 128) {
    final int value = in.read();
    for (int i = 0; i < (n & 0x7f); i++) {
     out[position++] = (byte) value; }
   } else {
    for (int i = 0; i < n; i++) {
    out[position++] = (byte) in.read();}
   }
  }
 }
```

(b) Patched source code

```java
private static void decompress
 (final InputStream in, final byte[] out)
 throws IOException {
  int position = 0;
  final int total = out.length;
  while (position < total) {
   final int n = in.read();
   if (n == 0) { // `<' modified to `=='
    throw new ImageReadException("Error decompressing
          RGBE file"); }
   if (n > 128) {
    final int value = in.read();
    for (int i = 0; i < (n & 0x7f); i++) {
     out[position++] = (byte) value; }
   } else {
    for (int i = 0; i < n; i++) {
    out[position++] = (byte) in.read();}
   }
  }
 }
```

(c) Artificial Vulnerability

CVE-2018-17201 causes an infinite loop that may lead to a DoS attack. The original fix adds a conditional exception, while the modified version changes the condition and reintroduces the vulnerability.

Setup

14 LLMs

(1) DeepSeekR1 Qwen 32B
(2) GPT3.5 Turbo
(3) GPT3.5 Turbo 1106
(4) GPT3.5 Turbo Instruct
(5) GPT4
(6) GPT4 0613
(7) GPT4 Turbo
(8) GPT4o
(9) GPT4o Mini
(10) LLaMA 3.1 70B Instruct
(11) LLaMA 3.3 70B Instruct
(12) Mistral 7B v0.2 Instruct
(13) Mistral 7B v0.3 Instruct
(14) Mistral 8×7B v0.1 Instruct

56 Vulnerabilities (15 real, 41 artificial)

(1) APACHE-COMMONS-001
(2) CVE-2013-5960
(3) CVE-2014-4172
(4) CVE-2016-10006
(5) CVE-2016-2162
(6) CVE-2016-6802
(7) CVE-2017-5662
(8) CVE-2018-1000089
(9) CVE-2018-1000531
(10) CVE-2018-1000850
(11) CVE-2018-1000854
(12) CVE-2018-11771
(13) CVE-2018-17201
(14) CVE-2019-12402
(15) HTTPCLIENT-1803

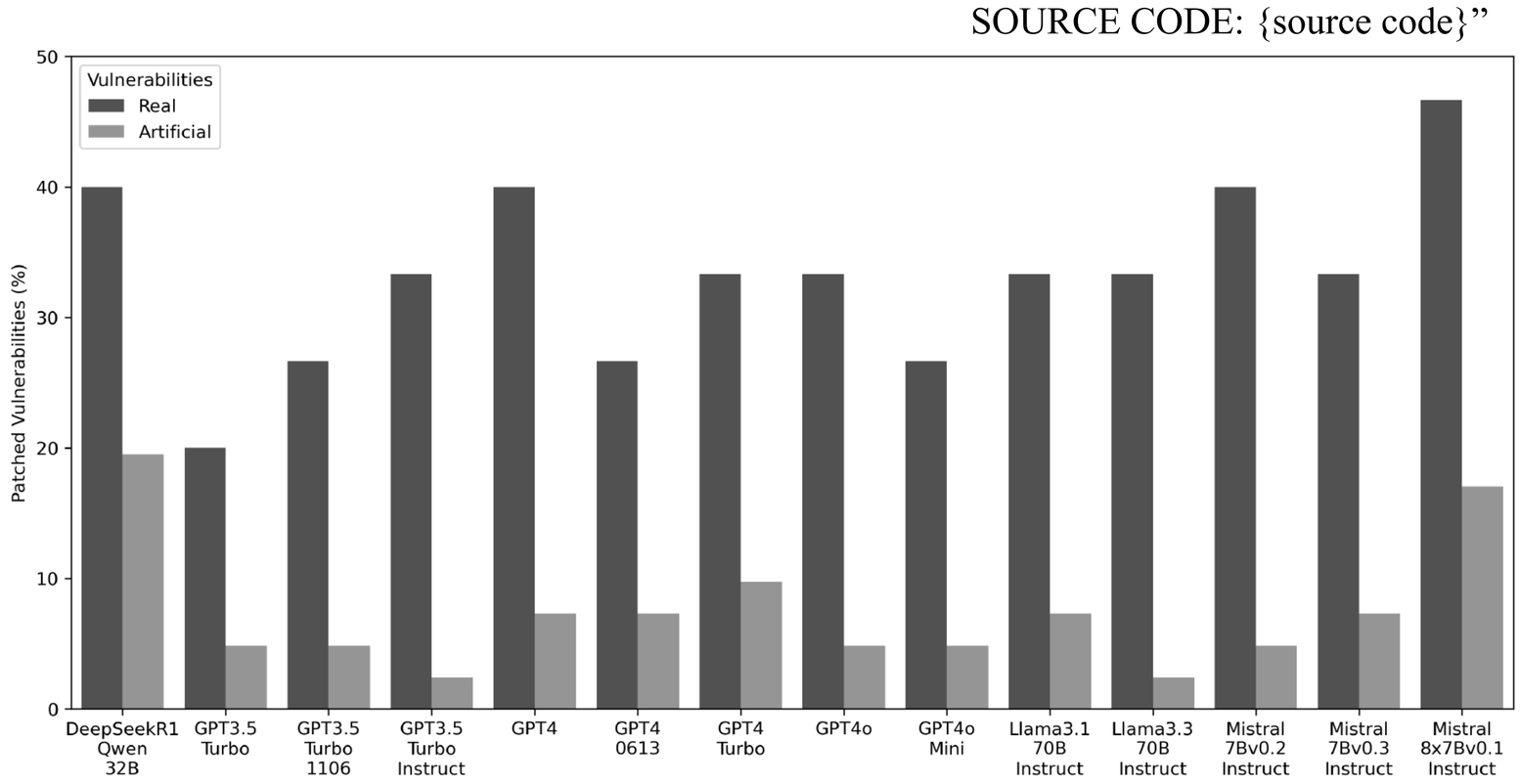
Prompting strategy

"You are a security expert who is good at static program analysis. The following SOURCE CODE, written in Java, contains a vulnerability. Please write the source code to fix this vulnerability. Do not make any other changes to the source code, and just reply with the final patched Java function. SOURCE CODE: {source code}"

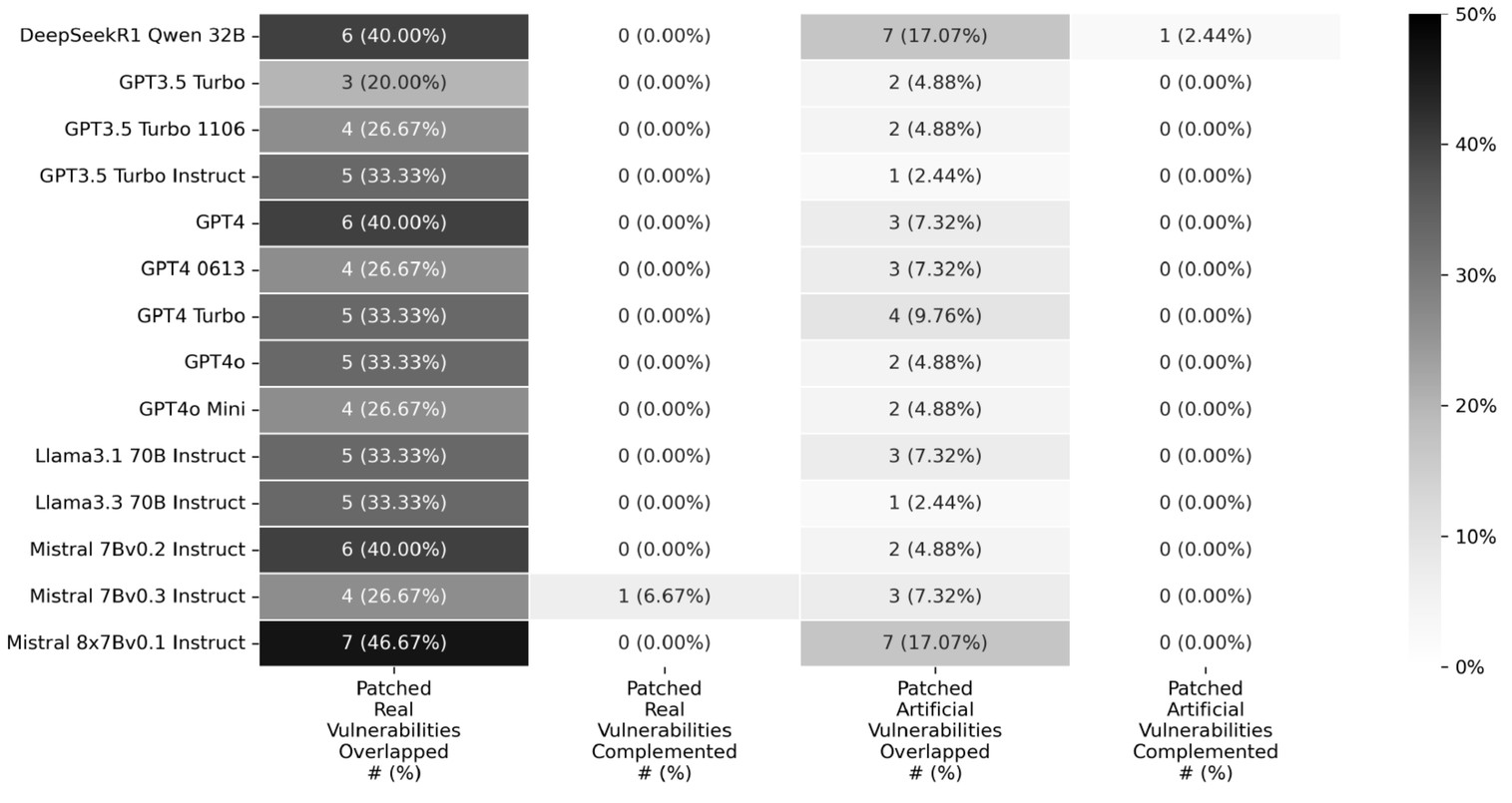
Results

RQ1 (Patching effectiveness of LLMs for real and artificial vulnerabilities)

- All evaluated LLMs patch real vulnerabilities more successfully than artificial vulnerabilities.
- This consistent gap across 14 models suggests that current one-shot LLM patching has limited generalization to artificial variants, even when those variants strongly couple to the corresponding real vulnerabilities.
- The strongest models, Mistral 8x7Bv0.1 Instruct and DeepSeekR1 Qwen 32B, each patch 14/56 vulnerabilities, however the performance still drops notably on artificial cases.



Patching effectiveness of LLMs for real and artificial vulnerabilities.



Overlapping and complementarity of LLMs in patching real and artificial vulnerabilities

RQ2 (Overlapping and complementarity of LLMs in patching real and artificial vulnerabilities)

- LLM overlap is much higher on real vulnerabilities than on artificial ones.
- Several LLMs converge on the same real vulnerabilities, however the overlapping drops sharply for artificial cases.
- At the same time, unique complementary patches are almost absent, with only two unique successes across all models and vulnerabilities.
- This indicates that combining current LLMs into ensembles is likely to provide only limited additional coverage, especially for the artificial cases.