This research paper introduces **ACH (Automated Compliance Hardener)**, a system developed at Meta for automated unit test generation, focusing on identifying and preventing regressions related to specific issues like privacy. Here's a breakdown of the key concepts and ideas:

## Core Idea: Mutation-Guided LLM-Based Test Generation

Instead of traditional methods, ACH uses a novel approach:

* **Targeted Mutation Testing:** ACH strategically introduces simulated faults ("mutants") specifically related to a chosen concern (e.g., privacy violations) into the code. These aren't random mutations; they're designed to mimic real-world bugs that could lead to the targeted issue.
* **LLM-Powered Test Generation:** ACH then uses a Large Language Model (LLM) to automatically generate new unit tests that can detect (or "kill") these mutants. The goal is to create tests that expose the potential vulnerabilities represented by the mutants.

In essence, ACH uses LLMs to create "super bugs" (mutants) and then generates tests to catch them.

## Key Aspects of ACH

* **Focus on Specific Concerns:** ACH isn't a general-purpose test generator. It's designed to harden code against particular issues, such as privacy, security, or integrity.
* **Agentic LLM Workflow:** ACH employs multiple LLM-based "agents" that work together to generate faults and tests.
* **Assurances:** ACH provides assurances about the tests it generates:
  + **Buildable:** The tests are syntactically correct and compile successfully.
  + **Valid Regression Tests:** The tests pass consistently on the original code.
  + **Hardening:** The tests catch faults that existing tests don't.
  + **Relevant:** The tests are closely related to the issue of concern (e.g., privacy).
  + **Fashion Following:** The tests adhere to the coding style of existing tests.
* **Mutation Testing Importance:** The paper emphasizes that mutation testing is superior to structural coverage criteria (like line coverage) because it can find faults even when the same lines of code are executed in a new way.
* **Wide Applicability:** While the paper focuses on privacy, the authors believe the ACH approach can be applied to various software testing problems.
* **Addressing a Fundamental Question:** ACH tries to bridge the gap between vague textual descriptions of software concerns and concrete unit tests that prevent bugs related to those concerns.
* **Risk Assessment:** ACH can use the distribution of simulated faults to estimate the risk exposure of different system components to the issue under test.
* **LLM Model:** The single language model Llama 3.1 70Bn was used in all the agents reported on.

## ACH System Architecture

The ACH system consists of several components working in a pipeline:

1. **Issue Summary:** Takes textual descriptions of concerns (e.g., previous privacy faults, regulatory requirements) as input.
2. **Fault Generation:** An LLM-based agent generates simulated faults (mutants) in the code, guided by the issue summary, existing tests, and the code under test.
3. **Equivalence Detection:** An LLM-based agent identifies and removes mutants that are functionally equivalent to the original code (i.e., they don't actually change the behavior).
4. **Test Generation:** An LLM-based agent generates new unit tests that are designed to "kill" the non-equivalent mutants.
5. **Testing and Validation:** The generated tests are built and executed against both the original code and the mutated code. Tests that pass on the original code but fail on the mutated code are considered valuable.
6. **Code Review:** Engineers review the generated tests and assess their relevance and quality.

## Key Contributions

* **Empirical Results:** The paper presents results from applying ACH to real-world software platforms at Meta, including the number of mutants generated, tests created, and the impact on code coverage.
* **Deployment Experience:** The paper shares insights from deploying and evaluating ACH at Meta, including engineer acceptance rates and feedback on the relevance of the generated tests.
* **Equivalent Mutant Detection:** The paper evaluates the performance of ACH's equivalent mutant detection agent.
* **Lessons Learned:** The paper discusses the challenges and open research questions that arose from this industrial application of LLM-based test generation.

## Prompts

The paper uses three simple prompts:

* **Make a fault:** Generates a mutated version of the class that contains a typical bug that introduces a privacy violation similar to a given diff.
* **Equivalence detector:** Determines if two versions of a Kotlin class will always do the same thing.
* **Make a test to catch fault:** Writes an extended version of the test class that contains extra test cases that will fail on the mutant version of the class but would pass on the correct version.

## Implications and Future Directions

This research suggests that mutation-guided LLM-based test generation is a promising approach for hardening software against specific concerns. The authors highlight the potential for further research in areas such as:

* More sophisticated prompting techniques for LLMs.
* Using language model ensembles to improve test generation.
* Applying the approach to a wider range of software testing problems.