

# An Agent-Based Framework for the Automatic Validation of Mathematical Optimization Models

Alexander Zadorojniy\*

IBM Research

[zalex@il.ibm.com](mailto:zalex@il.ibm.com)

Segev Wasserkrug

IBM Research

[segevw@il.ibm.com](mailto:segevw@il.ibm.com)

Eitan Farchi

IBM Research

[farchi@il.ibm.com](mailto:farchi@il.ibm.com)

## Abstract

Recently, using Large Language Models (LLMs) to generate optimization models from natural language descriptions has become increasingly popular. However, a major open question is how to validate that the generated models are correct and satisfy the requirements defined in the natural language description. In this work, we propose a novel agent-based method for automatic validation of optimization models that builds upon and extends methods from software testing to address optimization modeling. This method consists of several agents that initially generate a problem-level testing API, then generate tests utilizing this API, and, lastly, generate mutations specific to the optimization model (a well-known software testing technique assessing the fault detection power of the test suite). In this work, we detail this validation framework and show, through experiments, the high quality of validation provided by this agent ensemble in terms of the well-known software testing measure called mutation coverage.

## 1 Introduction

Mathematical optimization is a formal mathematical modeling technique that can be applied to help address many important real-world decision-making problems. Although our approach is applicable to general mathematical-programming problems, for clarity of exposition in this paper we focus on the subset of optimization problems that can be formulated as linear programs (LPs) of the following form:

$$\begin{aligned} & \text{minimize } c^T x && \text{(objective function)} \\ & \text{subject to } Ax \leq b, && \text{(inequality rows)} \\ & \ell \leq x \leq u && \text{(variable bounds)} \end{aligned}$$

Where  $x \in \mathbb{R}^n$  is a vector of decision variables,  $c \in \mathbb{R}^n$  is cost or profit vector,  $A \in \mathbb{R}^{m \times n}$ ,  $b \in \mathbb{R}^m$  is a matrix and right-hand side for the inequality block  $Ax \leq b$ ,  $\ell, u \in (\mathbb{R} \cup \{-\infty, \infty\})^n$  is component-wise lower and upper bounds; set  $\ell_i = -\infty$  or  $u_i = \infty$  when a bound is absent.

Modeling a real world problem using mathematical optimization is a lengthy process, requiring significant skills as it requires modeling a problem using a set of decision variables whose values should be selected so as to optimize some objective function while satisfying a set of constraints. Therefore, in recent years, many works have explored using LLMs to generate mathematical optimization models from natural language descriptions [1–4] so as to make optimization modeling more widely accessible and reduce the time required to create such models. Such efforts have met with considerable success, which has even resulted in some of the commercial software vendors offering LLM based *modeling assistants* [5]. However, in spite of this success there are often still mistakes in the LLM generated models. Therefore, a significant open question is how to automatically validate the correctness of such models.

---

\*IBM Research

This issue of generated model validation arises from two characteristics of optimization modeling: the first is the fact that the natural language description must be translated into a formal model, and the second is that many, significantly different model formulations can solve the problem specified in natural language. The reason that there is such a variety of ways to model the problem is that ultimately, two formulations are suitable for solving the same problem if they capture the same set of optimal solutions (a solution is optimal if it is a solution that maximizes/minimizes the value of the objective function while satisfying the constraints). To see a concrete example of such two different valid models, consider the following decision problem: A mid-sized precision-engineering plant manufactures two profitable items during each eight-hour shift - heavy-duty gearbox housings and precision mounting brackets. Every housing and every bracket requires one hour in total for deburring, gauging, and final sign-off at assembly and inspection benches. A housing occupies the machine and finishing cell for two hours, whereas a bracket needs one hour. The combined line is capped at ten machine-hours each day. Each gearbox housing results in 120\$ net profit per unit and each mounting bracket in 90\$ net profit per unit. The problem is to decide how many housing units ( $x$ ) and how many brackets ( $y$ ) to produce so as to optimize the profit while satisfying the time constraints of the plant. One possible formulation of the problem is:

$$\text{maximize } 120x + 90y \quad (1a)$$

$$\text{subject to } x + y \leq 8, \quad 2x + y \leq 10, \quad x, y \geq 0 \quad (1b)$$

An equivalent formulation of the problem is one in which the same set of constraints are used but the objective function is scaled to a different currency resulting in the objective **maximize**  $1200x + 900y$ . These two correct formulations will result in two different optimal objective function values, even though they will both provide the same optimal solutions in terms of the  $x$  and  $y$  values. Another equivalent optimization model is the *dual formulation* [6] which again results in an optimal solution to the original problem but with a very different formulation (equations (2a)-(2b) show the dual formulation for the problem described in equations (1a)-(1b)):

$$\text{minimize } 8u_1 + 10u_2 \quad (2a)$$

$$\text{subject to } u_1 + 2u_2 \geq 120, \quad u_1 + u_2 \geq 90, \quad u_1, u_2 \geq 0 \quad (2b)$$

Another reason for the use of different models for the same problem stems from the fact that once modeled in mathematical form, mathematical optimization models are often solved by generic **Optimization engines** such as IBM ILOG CPLEX Optimization Studio [7] or the Gurobi Optimizer [8], and different formulations could have a very significant impact on the time it takes such engines to find optimal, or even good, solutions. This results in additional incentives to create different, more efficient models, for the same problem.

Irrespective of the underlying reason, the existence of different equivalent models may mean that it is quite hard to understand whether any given model is indeed a model that provides an optimal or even a feasible solution for a business problem. This makes automatic validation of generated models especially challenging - even when a ground-truth model is available. Existing techniques for automatic model validation include testing syntactic equivalence between the generated and ground truth models [2] and comparing only their optimal objective values [1]. Both techniques are unreliable: both would fail, for example when two optimization models differ by the scaling of the objective function as described above, resulting in a false negative. The use of objective value comparison could also result in a false positive, as can be seen by considering the case illustrated in Figure 1c. This figure contrasts a correct formulation with an incorrect one that satisfies only a subset of feasible scenarios. If we evaluate the models on a default instance that the flawed formulation happens to satisfy, the comparison of objective values alone will yield a false positive. Consequently, relying exclusively on objective-value checks is insufficient for robust model validation.

**Contribution** To address this challenge, we propose a novel agent-based approach that builds upon best practices in software testing for validating optimization models. Our approach is an automated one, and includes a set of several LLM based agents that utilize the natural language description of the problem. These agents include: an agent that uses an LLM to create a problem level, rather than an optimization model level, testing interface - enabling the output of the model to be validated against the natural language specification of the problem in a model agnostic manner; an LLM agent to create unit tests - intended to validate model correctness by utilizing the solutions provided by the model through the business level interface; an LLM that generates an auxiliary optimization model aimed at providing feedback on the correctness of unit tests; and an LLM agent to generate optimization modeling specific *mutations* - a well-known software testing concept whose goal is to help assess the fault-detecting power of a test suite. We executed the test suite on the auxiliary optimization models deterministically, guaranteeing that every test yields a definitive Pass or Fail result. In the rest of this work we detail our approach, and demonstrate its effectiveness through empirical validation.

## 1.1 Related Work

**Optimization Modeling and Verification using LLMs** [3] propose a multi-agent LLM system that automatically derives optimization models from textual problem descriptions. Validation was carried out by a group of experts who wrote at least five tests per problem to verify code correctness. This strategy is labor-intensive, and a suite of only five tests is generally inadequate — even for academic benchmark datasets. For automatic validation, their workflow depends on an additional LLM-based agent to review model correctness; the agent’s feedback is then fed to the other agents to steer subsequent iterations. Because LLMs are susceptible to hallucination [9], this review loop is inherently error-prone.

[1] present an LLM-driven, multi-agent system that automatically constructs optimization models from natural-language problem descriptions. Although this is an important step toward fully automated model generation, their validation strategy compares only the optimal objective values of the generated models with manually computed reference values. As described in Section 1, agreement (resp. disagreement) on the objective value alone does not guarantee that a model is correct (resp. incorrect). Moreover, the true optimum is frequently unknown apriori — making this form of validation unreliable.

[4] proposes a system that answers users’ supply-chain queries (e.g., *what-if* analyses). The authors validated correctness by manually creating five scenarios with known optimal solutions. For each scenario, they automatically generated question-and-answer sets—for example, “What if we prohibit shipping from supplier X to roaster Y?” or “Demand at café Z increases by 10%.” Macros substitute random entity names, yielding thousands of distinct yet structurally similar questions. Although this approach provides valuable validation for the supply-chain-specific scenarios described in this work, crafting such templates for every problem is impractical, and optimal solutions are often unknown or not unique.

**Software Testing** Traditional software testing ensures that code behaves as expected by executing test cases that uncover defects, validate functionality, and exercise edge cases. Recently, large language models (LLMs) have been applied to automate and enhance this process. For example, Meta’s TestGen-LLM tool [10] automatically generates functional tests for conventional software.

Our work, by contrast, does not validate traditional software; instead, it targets the creation and verification of mathematical programming models. We nevertheless highlight recent advances in LLM-based test generation for completeness.

To improve test effectiveness and guarantee sufficient *test coverage*, a variety of coverage approaches have been proposed in the software testing domain [11]. One widely adopted coverage approach is mutation testing, which systematically creates small program variants (mutants) to assess the fault-detecting power of a test suite [12]. As a part of our work presented in this paper, we explore how the mutation-testing technique can be transferred to the domain of mathematical programming, using LLMs to automate mutant generation and evaluation.

## 1.2 Background - Mutation Testing

Mutation testing [13] is a technique used to evaluate the effectiveness of a test suite. It captures the intuitions that some categories of software bugs occur as a result of small perturbations to the software code, e.g., writing  $\leq$  instead of  $\geq$  or visa versa. Such changes to the code are called mutations. In more details the method works as follows.

1. **Mutations introduction:** Small changes, called mutants, are intentionally injected into the program's source code. These mutations might involve altering operators (e.g., changing  $+$  to  $-$ ), modifying constants, or tweaking control structures.
2. **Test Execution:** The entire test suite is executed against the program being tested and its mutants. The idea is to simulate potential faults that could occur in the code. Note that the fault could be in the original program as it should have had  $\geq$  in a condition but had  $\leq$  in a condition. Introducing a mutant that changes  $\geq$  to  $\leq$  is aimed at revealing that programming error.
3. **Tests effectiveness evaluation:** If a mutant causes the test suite to fail, it is considered "killed," indicating that the tests were effective in catching that change. If a mutant does not trigger any test failures, it "survives." A surviving mutant suggests that there might be a gap in the test suite, as it failed to detect a potential defect.
4. **Tests coverage improvement.** By identifying surviving mutants, developers can write additional tests or refine existing ones and ensure that the test suite is strong enough to catch subtle programming errors.

Common type of mutations include value and decision mutations. Value mutations involve changing the literal or constant values in the code. These mutations simulate scenarios where an incorrect value is used, helping ensure that the tests validate the proper handling of data. Decision mutations focus on the logical structure of the code, especially the conditions that control the flow of execution (such as if or while statements). These mutations simulate errors in decision-making logic.

## 1.3 Tests Coverage for Optimization Models

Software coverage [11] is aimed at giving a measure of the adequacy of testing. For example, it may count the percentage of statements that were executed by any test in the testing set. Compared to software development, optimization problems typically involve fewer programming constructs, often just a few dozen instead of thousands, and those requirements tend to be more clearly defined, change less frequently, and involve less integration with existing systems. However, optimization problems are more challenging when it comes to the mathematical modeling, the choice of algorithms and solvers, and managing computational complexity. Thus, we expect that testing and validation of optimization problems will be generally simpler than it is for large-scale software projects. For testing of the optimization models we will use decision and value mutations which should be sufficient to cover most frequent automatically generated optimization models issues.

**Mutation Coverage of Optimization Models** As described above, a common way of defining mutation coverage is through the ratio of number of "killed" mutations to overall mutations generated. In our context mutations are applied to the constraints of a given problem in the benchmark. For example, a constraint of a given problem in the benchmark may state that  $x + y \leq 8$  and its mutant may be  $x + y \leq 7$  (changing the constant 8 to the constant 7).

In addition, the concept of a *killed mutation* applies to our context. A mutant is said to be *killed* if we can demonstrate that the mutated mathematical program fails to pass the tests intended to validate that the model solves the optimization problem stated in natural language. Let  $K$  denote the number of

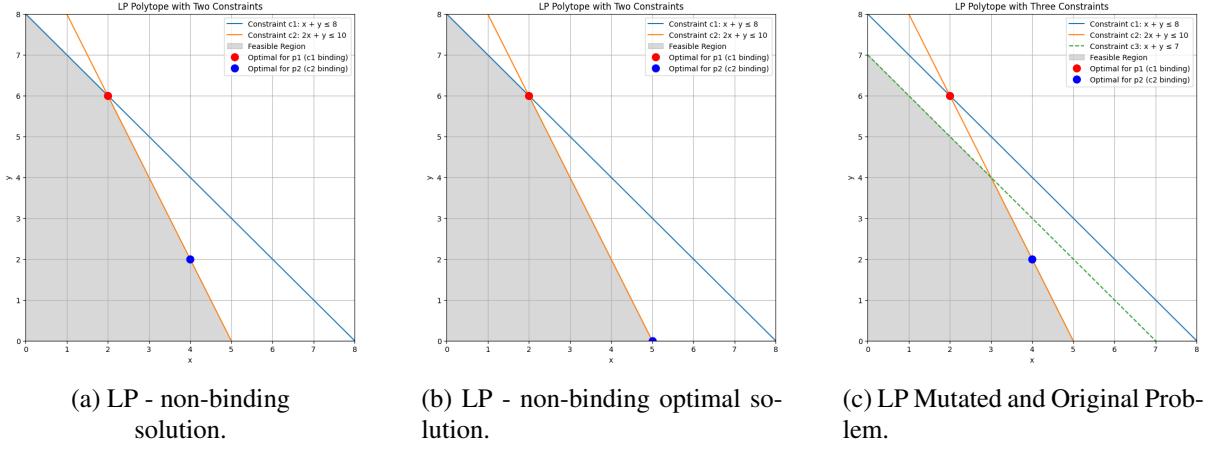


Figure 1: LP - What we want to cover.

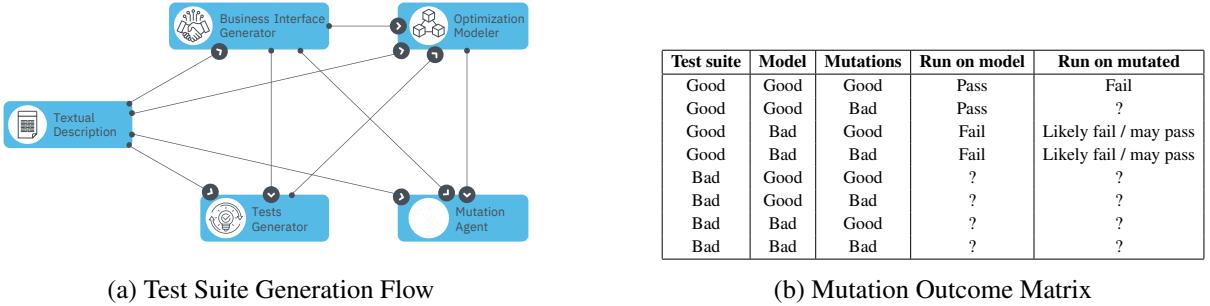


Figure 2: Flow and expected outcomes. (a) Test suite generation flow. (b) Outcome matrix for combining test-suite quality (Good/Bad), model correctness (Good/Bad), and mutation validity (Good/Bad): a good suite with a good model should pass on the base model and typically fail on a mutated model; “?” entries indicate cases where outcomes depend on specifics (e.g., weak suites or invalid mutations).

killed mutants for a dataset and  $M$  the total number of mutants for that dataset. The *mutation coverage* for the dataset (MC) is then defined as

$$MC [\%] = \frac{K}{M} \times 100 \%. \quad (3)$$

Throughout this paper, we define coverage as mutation coverage.

To illustrate the importance of mutation coverage, consider the same daily production-planning problem evaluated with two different profit functions but identical constraints. The original profit function is  $120x + 90y$  and the modified one is  $140x + 60y$ . With the correct set of constraints, both profit functions are supported—each yields an optimal, feasible solution (Figures 1a and 1b respectively). However, if constraint  $c1$  is mistakenly replaced by constraint  $c3$  (Figure 1c), only the second profit function remains supported.

This is precisely what mutation testing does: it deliberately alters the model to reveal potential errors.

## 2 Automatic Optimization Validation using LLMs

To close the gap between natural-language specifications and the robust validation of optimization solutions, we defined a framework composed of a workflow of four agents. The workflow (appearing in Figure 2a) begins with a *Textual Description* authored by a domain expert.

This description is parsed by the **Business-Interface Generator**, an agent which converts the natural language problem statement into a unified, declarative interface that defines the form of the solution

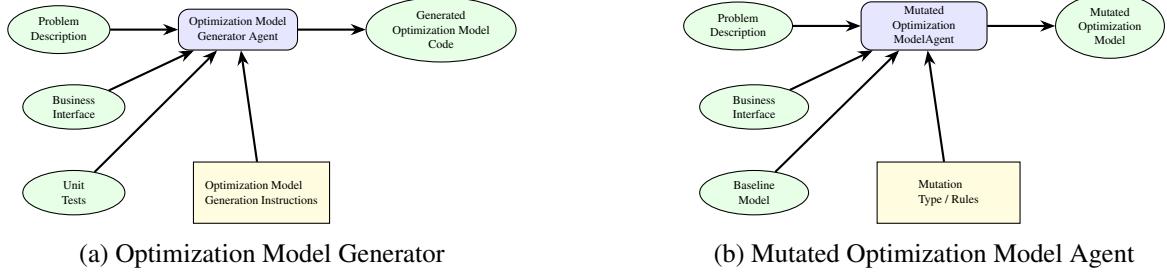


Figure 4: Agents and their I/O. (a) Optimization Model Generator — inputs: problem description, business interface, unit tests, and generation instructions; output: optimization model code. (b) Mutated Optimization Model Agent — inputs: problem description, business interface, baseline model, and mutation rules; output: mutated optimization model.

that needs to be returned by the solution of the optimization model. All downstream components can consume this interface, which defines entities, parameters, and KPIs related to the solution.

Next, the **Tests Generator** agent uses both the interface and the natural-language specifications to iteratively assemble a diverse suite of test instances, that characterizes the expected behavior of any subsequent optimization artifact.

The **Optimization Modeler** agent then builds an auxiliary optimization model that returns an optimization solution conforming to this interface (variables, constraints, and objective) which is used to verify the validity of the tests produced by the tests generator.

To ensure that the tests generated by the generator provide good testing coverage, a dedicated **Mutation Agent** creates target mutations which can be injected into the optimization model, so that the tests can be rerun on the mutated models.

Surviving mutants indicate potential weaknesses in the test suites, whereas killed mutants strengthen confidence in both the correctness of the test suite and the generated optimization model. For each agent we provide a diagram which describes the flow for that agent (Figures 3a, 3b, 4a, 4b, 5b)

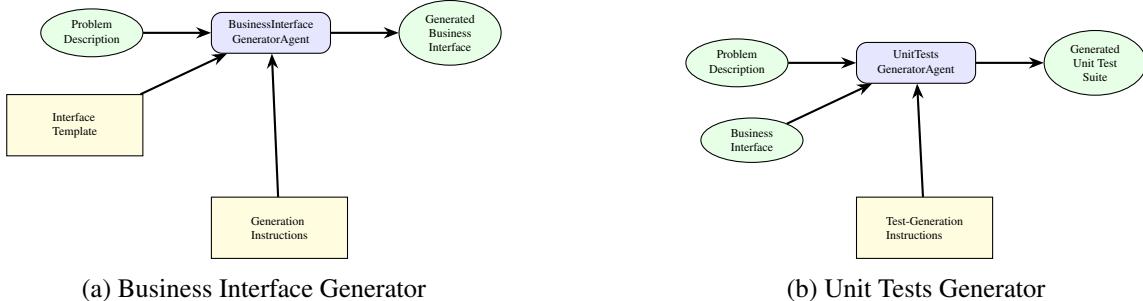


Figure 3: Agents and their I/O. (a) Business Interface Generator — inputs: problem description, interface template, and generation instructions; output: business-interface code. (b) Unit Tests Generator — inputs: problem description, business interface, and test-generation instructions; output: unit-test suite.

Since LLMs are prone to hallucination [9], and as we have the agents for the business interface, test suite generation, and model, a single run of the test-suite generation flow may not produce a high-quality outcome, due to errors either one or more of these (to see this consider Table 2b). It is clear that for generating a good test suite, our goal is to be in the first row of the table - the one in which the originally generated model passes the test suite but the mutated model does not; therefore, we run the flow until the generated model passes the test suite (or up to a maximum number of iterations). Although we do not establish a theoretical upper bound on the required number of iterations, we empirically show that it remains small on the well-known comprehensive optimization model dataset [14]. This iterative execution of the workflow appears in Algorithm 1. Note that as the model is not provided as input to the

test generator the test suite generation does not depend on the model produced.

---

**Algorithm 1** Tests Suite Generation Algorithm.

---

**Require:** Textual description  $TD$   
**Ensure:** Business interface  $BI$ , test suite  $TS$ , optimization model  $OM$ , mutation artifacts  $MA$

```

/* Initialization from the specifications */
1:  $BI \leftarrow \text{BUSINESSINTERFACEGENERATOR}(TD)$ 
/* Iterative refinement */
2: repeat
3:    $TS \leftarrow \text{TESTSGENERATOR}(TD, BI)$ 
4:    $OM \leftarrow \text{OPTIMIZATIONMODELER}(TD, BI, TS)$ 
5:   Run  $TS$  for  $OM$ 
6:    $MA \leftarrow \text{MUTATIONAGENT}(TD, BI, TS, OM)$ 
7: until  $TS$  and  $OM$  either Pass or Number of Iterations Outed           ▷  $MA$  is a byproduct
8: return  $(BI, TS, OM, MA)$ 

```

---

**Algorithm 1** begins with a natural-language specification  $TD$  and first extracts a high-level business interface  $BI$  (line 1). It then enters an iterative loop (line 2) in which (line 3) a test suite  $TS$  is re-generated from both  $TD$  and the current  $BI$ , (line 4) an optimization model  $OM$  is (re)constructed to satisfy  $TD$ ,  $BI$ , and the new tests, (line 5) the tests are executed against  $OM$ , and (line 6) a mutation agent produces artifacts  $MA$  that stress-test the system. The loop terminates when the model passes all tests (or a number of iterations budget elapses), after which the algorithm returns the four artifacts  $(BI, TS, OM, MA)$ , with  $MA$  serving as a useful by-product for future robustness checks.

Note that Algorithm 1 operates similarly to a "Monte Carlo Search", as it does not use feedback from failures in previous iterations. This is based on our observation that such feedback does not improve LLM performance for the experiments we did; however, it may still prove useful for more complex problems, in which case it could be incorporated as an automatic prompt-adjustment mechanism for the agents.

### 3 Applying Existing Test Suites to Legacy Optimization Models

The test-suite generator produces an auxiliary optimization model. Ultimately, however, we want to enable users to test optimization models that were created outside of this framework. To accommodate this, we developed an additional agent that adapts the generated test suite to any existing optimization model and executes the tests against it.



Figure 5: Agents and their I/O. (a) Tests Adjuster — input: problem description, optimization model, and original unit tests; output: adjusted API-aligned test suite. (b) Test Adjuster Agent — same I/O represented as a process diagram.

The figure illustrates a streamlined *test-adjustment workflow* that utilized this agent. On the left

reside the three artifacts available: the natural-language specification (**TD**), the external optimization model (**OM**) to be tested, and a regression test-suite (**Tests**). These artifacts are simultaneously fed into the **Test Adjuster** agent, whose task is to reconcile the suite with both the narrative description and the mismatched assertions related to API, updating parameters and their names. The resulting, consistently updated suite is executed against the optimization model.

---

**Algorithm 2** LLM based Adjustments of the Test Suite.

---

**Require:** Textual description  $TD$ , Optimization Model  $OM$ , Tests.

$AT \leftarrow \text{TESTSADJUSTER}(TD, OM, Tests)$

2: **return**  $AT$

---

**Algorithm 2** receives the textual specification  $TD$ , optimization model  $OM$ , and a pre-existing test suite. It invokes a large-language-model agent, *TestsAdjuster*, which cross-checks  $TD$  against the behavior of  $OM$  and revises the tests accordingly, producing a new artifact  $AT$ . The adjusted test suite  $AT$  is returned for subsequent validation cycles, ensuring that the evolving model remains aligned with both the specification and the optimization model.

## 4 Experiments

**Data** To assess the proposed approach, we drew a random sample of 100 problems from the NLP4LP benchmark (three problems —specifically 89, 124, and 269— were excluded because their descriptions were ambiguous).<sup>1</sup>

The NLP4LP data set consists of problems described as text (e.g., a basic scheduling or resource allocation task), constraints (e.g., limits such as capacity or budget constraints) and a goal (e.g., minimizing cost or maximizing profit). Each problem’s folder is organized in a modular way so that each problem instance is packaged with several files that separate its different aspects. For our experimentation we used `description.txt` for problems description and `solution.json` for solution comparison for default scenarios.

**Computational Experiments - Tests Suite Quality Validation** The tests for each problem were generated using our workflow with two different large-language-model (LLM) configurations:

- all agents using the `o1-preview` LLM; and
- a hybrid setup that combines `gpt-4o` for the auxiliary optimization model generator, and `o1-preview` for all other agents (the goal of this was to test the impact of a less powerful auxiliary optimization model generator on the validity of the generated test suite).

In both cases, the temperature was fixed at  $T = 1.0$ . In the hybrid configuration, `gpt-4o` LLM was used exclusively for the OM agent. Since a temperature of  $T = 1.0$  introduces non-deterministic behavior in the LLMs, we ran<sup>2</sup> each configuration twice for each problem in the selected dataset, and the entire experimental study took several hours to complete.<sup>3</sup>

Table 6a summarizes the effectiveness of the mutation-testing campaign for two language-model configurations. Across both the weaker (joint `gpt-4o` and `o1-preview`) and stronger (`o1-preview` for all agents) configuration, mutation coverage remained high—at least 69%—despite injecting only a single mutation per problem. Note that a single mutation is designed to mimic the most difficult and frequent type of error to detect. As expected, using a stronger LLM for the auxiliary optimization model

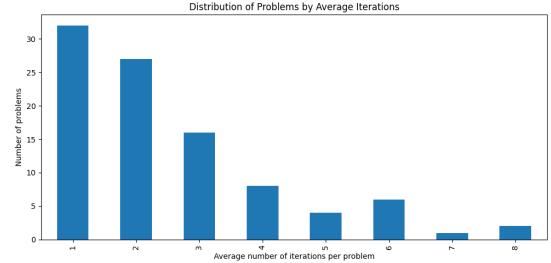
<sup>1</sup>All problem instances, source code, and prompts are provided in the supplementary material. The prompts were tuned using problems that are not part of the selected set of 100 problems.

<sup>2</sup>LLM access was provided via the Azure OpenAI service. Pre- and post-processing were performed on a machine with an 11th Gen Intel(R) Core(TM) i7-11850H CPU @ 2.50GHz and 64GB RAM.

<sup>3</sup>All models, test suites, and their interpretations are provided in the supplementary material.

Model	KILLED	NO	Ratio
all o1-preview	142	46	0.76
o1-preview + gpt-4o	131	58	0.69

(a) Mutation Kill Ratios



(b) Iteration Distribution

Figure 6: Mutation analysis and convergence. (a) Mutation kill ratios across models: `o1-preview` achieves a higher ratio (0.76) compared to the combined `o1-preview + gpt-4o` (0.69). (b) Distribution of the average number of iterations required for convergence, showing most problems converge within a few iterations, with a heavy right-skewed tail.

generator resulted in a higher quality set of tests as evidenced by the higher kill ration (76% as oppose to 69%).

Another important characteristic evaluated in our framework is the number of iterations required to converge to a “good” test suite. Figure 6b presents a histogram of the *average iterations-to-convergence* recorded for each optimization problem in the benchmark. The distribution is clearly right-skewed.

The vast majority of instances converge quickly: over three-quarters ( $\approx 76\%$ ) of the problems require no more than 3.5 iterations on average, with the most frequent class centered at 2.0 iterations—corresponding to approximately 18 problems. A diminishing tail extends to higher iteration counts. Only a small fraction (around 10%) of problems require five or more iterations, and the most demanding outlier averages 8.5 iterations.

Across both configurations, the `gpt-4o` variant required, on average, just under three iterations (2.8) to converge, whereas the `o1-preview` variant converged in approximately 2.5 iterations. To assess the statistical properties of the difference in iterations per problem, we applied the Shapiro–Wilk [15] test to the vector of pairwise differences between configurations. The resulting p-value was very small ( $2.17 \times 10^{-6}$ ), indicating that the distribution of differences deviates significantly from normality. Consequently, we employed the Wilcoxon signed-rank test [16] to evaluate whether the difference between the two configurations is statistically significant. The test yielded a p-value of approximately 0.4, suggesting that the observed difference is not statistically significant.

**Auxiliary Optimization Model Validation** As can be seen from the high kill ratio, most auxiliary optimization models were classified by our system as correct. This outcome supports the idea that small changes to a correct model are effectively detected by our mutation process. To further validate these results, we also tested the correctness of the auxiliary models created by the optimization generator agent using an alternative method. Specifically, we evaluated whether each generated model, when provided with the input parameters of its corresponding problem instance from the optimization dataset, produced the correct objective function value. This evaluation was conducted across all models generated in all iterations. If the correct reference value was reproduced in the final iteration, we labeled the auxiliary model for that problem as **correct**.

For 66 problems, the reference value was matched in all four runs. For another 12 problems, the reference value was matched in exactly three runs, which led on average to nine auxiliary models per problem being classified as correct. For the remaining cases where the reference value was reproduced in only 0, 1, or 2 runs, we manually inspected both the auxiliary models and the published reference solutions. Based on this inspection, we identified, on average, 12 additional auxiliary models as correct. In total, 87/97 auxiliary models were deemed correct, corresponding to approximately 90% accuracy in the auxiliary optimization models generated. This further attests to the validity of the test suites generated by our framework, which consistently classified original optimization models as valid and

their mutated variants as invalid with a high kill ratio.

#### 4.1 Computational Experiments - Testing external optimization models

To evaluate the usefulness of our testing framework on optimization models generated externally, we required models outside our own pipeline. The dataset we used included only Gurobi solver Python code generated by the Optimus system [1], which was not suitable for our purposes, as we did not use the Gurobi solver (we used the CPLEX solver instead). In our work, we relied on the Pyomo Python package, as it supports multiple solvers rather than being limited to a specific one.

We therefore evaluated the testing of external models on nine problems from the dataset using the following procedure: we ran the test suites produced by our workflow using only the `o1-preview` model on optimization models generated for the same problems by the joint `o1-preview + gpt-4o` pipeline. These models served as external inputs to the end-to-end `o1-preview` workflow. We selected nine problems (9, 20, 49, 95, 99, 105, 158, 191, and 199) for which the reference solutions in the NLP4LP dataset differed from those produced by the joint `o1-preview + gpt-4o` models. These cases were particularly valuable for cross-checking, as the test suites from the two workflows (`o1-preview` alone vs. joint `o1-preview + gpt-4o`) produced different results with respect to the reference solutions in the NLP4LP dataset. Despite mismatched interfaces between models and across iterations (although interfaces remained consistent within a single LLM model and iteration), the adjuster component successfully adjusted every test suite API. This allowed each model generated by `gpt-4o` to be executed and evaluated against its intended requirements using the test suite generated by the `o1-preview` model.

Table 1: Tests suite results generated by all *o1 – preview* LLMs for combination of *o1 – preview* and *gpt – 4o* optimization models.

Problem	Result	Description
9	FAIL	Correct optimization model
20	FAIL	Wrong optimization model
49	PASS	Correct optimization model
95	PASS	Correct optimization model
99	FAIL	Wrong optimization model
105	PASS	Correct optimization model
158	PASS	Correct optimization model
191	FAIL	Correct optimization model
199	PASS	Correct optimization model

We then examined each of the nine benchmark problems listed in Table 1 manually, and carried out additional targeted LLM checks, to determine whether its optimization model is *Correct* or *Incorrect*. Seven models were found to be correct, while two (i.e., problem 99 and problem 20) exhibited errors: one inequality constraint is written in the wrong direction, which can make the model infeasible or drive it to a sub-optimal solution, in another one the decision variables are declared as non-negative reals, but they should be integers; the domain is therefore too relaxed. To understand the results, refer again to Table 2b . As to generate the tests we used the iterative procedure outlined in Section 2, we assume that the tests are good with a high likelihood (i.e., the desired results are in rows 3 and 4 of the table). We expect good models to pass the tests and bad models to fail. In the empirical results, the test suite behaved as follows:

**Failed** (as desired) on Problems 20 and 99; **Passed** on Problems 49, 95, 105, 158, and 199 (all correctly modeled, as desired); **Failed** on Problems 9 and 191 despite correct models—these are false positives.

To summarize, our test suite correctly classified 5 models as correct and 2 models as incorrect. No optimization models were falsely classified as correct, while 2 models were falsely classified as incorrect—overall demonstrating good accuracy in this test.

## 5 Conclusions and Further Work.

We introduced an automated, multi-agent framework for validating optimization models that are generated by LLMs from natural-language descriptions. Borrowing techniques from software testing, the framework (i) creates a problem-specific testing API, (ii) automatically composes test cases using this API, and (iii) applies optimization-oriented mutation testing to probe edge cases. Experiments show that the approach achieves strong mutation-coverage scores, demonstrating its effectiveness at uncovering errors and confirming model correctness. As future work, two main directions are of interest: (i) refining the mutation process to achieve higher coverage, and (ii) evaluating the framework on more complex, real-world problems.

## References

- [1] Ali AhmadiTeshnizi, Wenzhi Gao, and Madeleine Udell. Optimus: scalable optimization modeling with (mi)lp solvers and large language models. In *Proceedings of the 41st International Conference on Machine Learning*, ICML’24. JMLR.org, 2024.
- [2] Rindranirina Ramamonjison, Timothy Yu, Raymond Li, Haley Li, Giuseppe Carenini, Bissan Ghaddar, Shiqi He, Mahdi Mostajabdeh, Amin Banitalebi-Dehkordi, Zirui Zhou, et al. Nl4opt competition: Formulating optimization problems based on their natural language descriptions. In *NeurIPS 2022 Competition Track*, pages 189–203. PMLR, 2023.
- [3] Ziyang Xiao, Dongxiang Zhang, Yangjun Wu, Lilin Xu, Yuan Jessica Wang, Xiongwei Han, Xiaojin Fu, Tao Zhong, Jia Zeng, Mingli Song, and Gang Chen. Chain-of-experts: When LLMs meet complex operations research problems. In *The Twelfth International Conference on Learning Representations*, 2024.
- [4] Beibin Li, Konstantina Mellou, Bo Zhang, Jeevan Pathuri, and Ishai Menache. Large language models for supply chain optimization, July 2023.
- [5] Dan Steffy. Introducing gurobi ai modeling, November 2024. Accessed: 2025-05-26.
- [6] Dimitris Bertsimas and John N Tsitsiklis. *Introduction to linear optimization*, volume 6. Athena scientific Belmont, MA, 1997.
- [7] IBM Corporation, Armonk, NY. *IBM ILOG CPLEX Optimization Studio 22.1.2*, version 22.1.2 edition, 2025. Accessed 12 July 2025.
- [8] Gurobi Optimization, LLC, Beaverton, OR. *Gurobi Optimizer Reference Manual*, version 12.0 edition, 2025. Accessed 12 July 2025.
- [9] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2), 2025.
- [10] Nadia Alshahwan, Jubin Chheda, Anastasia Finogenova, Beliz Gokkaya, Mark Harman, Inna Harper, Alexandru Marginean, Shubho Sengupta, and Eddy Wang. Automated unit test improvement using large language models at meta. In *Companion Proceedings of the 32nd ACM International Conference on the Foundations of Software Engineering*, pages 185–196, 2024.
- [11] Qian Yang, J Jenny Li, and David Weiss. A survey of coverage based testing tools. In *Proceedings of the 2006 international workshop on Automation of software test*, pages 99–103, 2006.
- [12] Soukaina Hamimoune and Bouchaib Falah. Mutation testing techniques: A comparative study. In *2016 international conference on engineering & MIS (ICEMIS)*, pages 1–9. IEEE, 2016.

- [13] Mike Papadakis, Marinos Kintis, Jie Zhang, Yue Jia, Yves Le Traon, and Mark Harman. Chapter six - mutation testing advances: An analysis and survey. volume 112 of *Advances in Computers*, pages 275–378. Elsevier, 2019.
- [14] Ali AhmadiTeshnizi, Wenzhi Gao, and Madeleine Udell. Nlp4lp. <https://huggingface.co/datasets/udell-lab/NLP4LP>, 2024. Version 1.0, CC BY-NC-SA 4.0. Accessed 14 Jul 2025.
- [15] Samuel Sanford Shapiro and Martin B Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3-4):591–611, 1965.
- [16] William Jay Conover. *Practical nonparametric statistics*. John Wiley & Sons, 1999.