LLM-based Coverage Analysis of Tests and Requirements

Francesco Mariotti, Lorenzo Sarti, Stefano Bacherini, Iacopo Trotta
Alstom Ferroviaria S.p.A.

Abstract

The correct management of traceability and coverage between tests and requirements can stem the late discovery of issues, which could have significant impact on projects. LLM capabilities can be used to perform a semantic comparison between the text of requirements and tests, to classify if a test is covering or not a specific requirement. In this work, an LLM-based approach for the analysis of coverage between tests and requirements is shown and evaluated through classification metrics.

Keywords: railway signaling; RAMS, V&V; safety demonstration.

1 Introduction

Railway control and protection systems are safety critical and they are regulated by specific standards such as EN 50716, EN 50126, and EN 50129. These emphasize rigorous verification and validation (V&V), systematic requirement engineering, and the construction of a safety case: therefore, requirements to test traceability is essential because it is part of the safety demonstration argument. Proper management of test and requirement traceability can avoid the late finding of problems with significant cost and time reduction [1]. However, such management is often done in peer-review through a Requirement Validation Table (RVT), which can lead to systematic errors.

Large Language Models (LLMs) offer powerful capabilities for semantic comparison, contextual interpretation, and similarity analysis of natural-language artifacts [2][3]: LLMs can make evident latent or non-obvious relationships that traditional traceability tools may overlook [4] with an automated semantic matching that provides analysts with timely and actionable feedback on the completeness and correctness of traceability links.

In this work, we leverage these capabilities to automatically identify potential correspondences between test descriptions and requirements, with a traditional quantitative evaluation to determine the reliability of LLM-based semantic traceability.

2 Methodology

The idea (implemented through a Python pipeline) can be schematized as follows:

1. RVT is the input, the information therein consists of the test ID and descriptions, the requirements and the associated tests.

2. The prompts to be provided to the LLM are created: one single prompt is built for each requirement/test couple: by providing the description of both, the prompt is written to understand if the given test is totally, partially or not covering the given requirement.
3. The prompts are launched through API calls to the LLM.
4. The results are gathered and shown to the user.

2.1 Evaluation metrics

To evaluate the approach, the ground truth are the RVTs already validated in former projects. As the approach addresses a classification problem, the classic confusion matrix is considered [5], as shown in Table 1. Classic evaluation metrics such as Precision and Recall are not relevant in this case, due to the high occurrence of TNs (given a test, it is natural that it will cover only a subset of requirements). Instead, the most relevant metric is the percentage of FNs: an FN means a missing test/requirement traceability, possibly leading to late discovery of issues.

2.2 Evaluation campaigns

The following evaluation campaigns were conducted to assess the effectiveness of the approach:

- **EC1** - consistency on repetition: experiments are replicated with the same setting multiple times (trade-off between statistical relevance and cost).
- **EC2** - prompt impact: using a defined project, change the prompt to possibly improve the outcome.
- **EC3** – performance in different projects: run the pipeline on different projects to evaluate how the system performs with different requirements and tests.

3 Results and discussion

EC1 was conducted on Project1, having 66 requirements and 34 tests, hence resulting in 2244 prompts. The results are shown in Table 2. It is possible to notice that the percentage of FNs is low with respect to the population. Moreover, the standard deviation is low, meaning that the results obtained from the LLM are stable with respect to repetitions.

For EC2 we still used Project1, using three different prompts with an increasing degree of details: Prompt1 containing just the instruction and the description of requirement and test; Prompt2 providing also some definitions, e.g., test coverage; Prompt3 containing also some examples of test/requirement coverage. The results are shown in Table 3. Differences between Prompt1 and Prompt2 are negligible, while for Prompt3 it is possible to notice a slight decrease in the percentage of FNs, but with a higher execution time.

Finally, in EC3 we experimented with different projects coming from main line and urban domains. Results in Table 4 show that the percentage of FNs across the different projects are compatible, meaning that the approach can be applied to different sectors.

4 Final Remarks

In this work we presented an LLM-based approach for the identification of tests/requirements coverage, which can be used as a support tool for the analyst to avoid the late detection of issues. The evaluation campaigns showed that the approach is consistent and applicable to different domains.

References

[1] M. Ruiz, J.Y. Hu, F. Dalpiaz (2023). Why don't we trace? A study on the barriers to software traceability in practice. Requirements Eng 28.

[2] Y. Zhu, J.R. Antony Moniz, S. Bhargava, J. Lu, D. Piraviperumal, S. Li, Y. Zhang, H. Yu, and B. Tseng (2024). Can Large Language Models Understand Context? In Findings of the Association for Computational Linguistics: EACL 2024, St. Julian's, Malta. Association for Computational Linguistics.

[3] E.E. Erkan, M.A. Kömürcü, T. Çelikten, A.E. Ergün, A. Onan (2025). Understanding Large Language Model Performance in Question Answering: A Comparative Analysis of Semantic and Lexical Metrics. In: Kahraman, C., et al. Intelligent and Fuzzy Systems. INFUS 2025. Lecture Notes in Networks and Systems, vol 1529. Springer, Cham.

[4] S.J. Ali, V. Naganathan, D. Bork (2025). Establishing Traceability Between Natural Language Requirements and Software Artifacts by Combining RAG and LLMs. In: Maass, W., Han, H., Yasar, H., Multari, N. (eds) Conceptual Modeling. ER 2024. Lecture Notes in Computer Science, vol 15238. Springer, Cham.

[5] S.V. Stehman (1997), Selecting and interpreting measures of thematic classification accuracy, Remote Sensing of Environment, Volume 62, Issue 1.

Table 1 - Confusion Matrix

		Actual Values	
		Positive	**Negative**
Predicted Values	**Positive**	**True Positive (TP):** test indicated by the LLM as covering the requirement and present in the RVT	**False Positive (FP):** test indicated by the LLM as covering the requirement but not present in the RVT
	Negativ	**False Negative (FN):** test not indicated by the LLM as covering the requirement but present in the RVT	**True Negative (TN):** test not indicated by the LLM as covering the requirement and not present in the RVT

Table 2 - Results for EC1

Experiment ID	TP	TN	FP	FN	Accuracy	FN %
EC1_01	65	1974	92	113	0,909	5,036
EC1_02	68	1971	95	110	0,909	4,902
EC1_03	66	1976	90	112	0,91	4,991
EC1_04	66	1982	84	112	0,913	4,991
EC1_05	65	1973	93	113	0,908	5,036
EC1_06	63	1977	89	115	0,909	5,125
EC1_07	63	1972	94	115	0,907	5,125
EC1_08	73	1979	87	105	0,914	4,679
EC1_09	67	1981	85	111	0,913	4,947
EC1_10	67	1967	99	111	0,906	4,947
Standard Deviation	2,722	4,467	4,467	2,722	0,003	0,017

Table 3 - Results for EC2

Experiment ID	Prompt ID	Execution Time	TP	TN	FP	FN	FN %
EC2_01	Prompt1	08m31s	73	1979	87	105	4,679
EC2_02	Prompt1	08m37s	67	1981	85	111	4,947
EC2_03	Prompt1	07m45s	67	1967	99	111	4,947
EC2_04	Prompt2	08m02s	66	1984	82	112	4,991
EC2_05	Prompt2	08m15s	68	1976	90	110	4,902
EC2_06	Prompt2	08m29s	70	1974	92	108	4,813
EC2_07	Prompt3	33m53s	103	1658	408	75	3,342
EC2_08	Prompt3	38m35s	108	1666	400	70	3,119
EC2_09	Prompt3	35m25s	109	1677	389	69	3,075

Table 4 - Results for EC3

Experiment ID	TP	TN	FP	FN	Accuracy	FN %
EC3_PJ1	66	1984	82	112	0,914	4,991
EC3_PJ2	25	248	63	16	0,776	4,545
EC3_PJ3	73	873	219	55	0,775	4,508
EC3_PJ4	32	775	8	10	0,978	1,212
C3_PJ5	23	781	16	16	0,962	1,914