Evaluating Template- and Model-Guided LLMs for Requirements Specification

Ikram Darif, Zainab Benlagote, and Ghizlane El Boussaidi

University of Ottawa, Ottawa, Canada
École de Technologie Supérieure, Montréal, Canada

MODRE - RE 2026, August 17, 2026
Montreal, Quebec, Canada

Research context

Specified using

Requirements

Templates

If <condition>, the <system name> shall <action> <object>

↓

If the user clicks on the create button, the website shall create the account

- × Resource-intensive process.
- × Extensive manual labour.
- × Requires domain expertise.

- ✓ Improves quality while preserving readability of NL requirements.
- ✓ Enables the automated analysis of requirements.
- ✓ Supports the verification and certification of safety critical systems.

Existing Work Limitations

1

Existing approaches still require substantial manual effort for template completion and correctness verification.

2

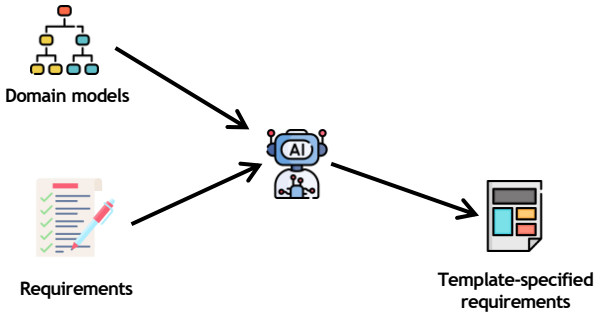
Existing approaches rarely exploit MDE, despite its demonstrated benefits for downstream RE tasks.

3

No existing approach combines the reasoning capabilities of LLMs with MDE to support template-based NL requirements specification.

Research Objectives

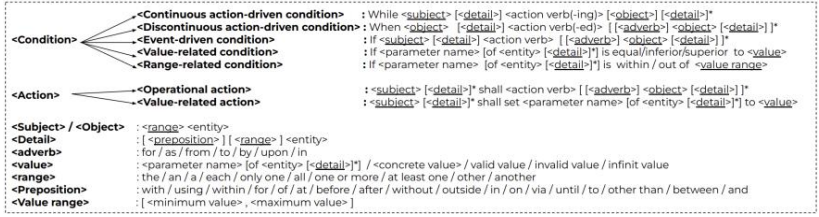
- Empirically evaluate the use of LLMs for generating template-based requirement specification.
- Examine whether domain models provide greater benefits than unstructured textual context when enriching prompts.



Research Background

- In this empirical evaluation, we specifically focus on the operational behavior template, a functional requirement template developed in our previous work (RESPECT) *[1]*, guided by the ARINC-653 certification standard *[2]*.
- The Operational behavior template was selected as it was the most frequently used template in our previous work *[1]*.

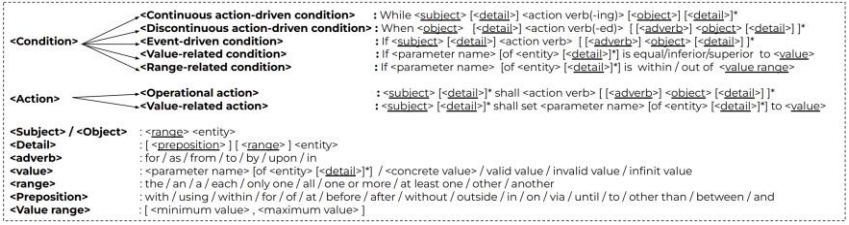
[<condition>
[AND <condition>]* (one per line)
Then]
<operational behavior action>



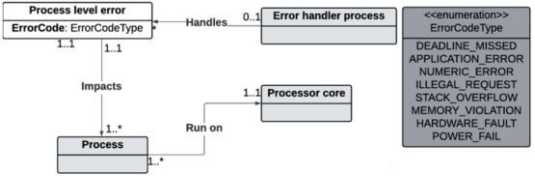
When **the OS** **calls** the **service**
AND
If **ERROR_CODE** is equal to **DEADLINE_MISSED**
Then
The **OS** shall set **RETURN_CODE** to **INVALID_PARAM**

Research Background

- Our previous work, RESPECT, supported the automated verification of requirements using domain models.



When **the OS** **calls** the **service**
AND
If **ERROR_CODE** is equal to **DEADLINE_MISSED**
Then
The **OS** shall set **RETURN_CODE** to **INVALID_PARAM**



Research Method

Research Questions

RQ.1:
To what extent do SOTA LLMs and different prompting strategies effectively generate template-based requirements specifications?

RQ.2:
How does the use of domain models as RAG context compare to unstructured textual context for template-based requirements specifications?

Datasets

- The empirical evaluation covers two services from the ARINC-653 certification *[2]*.
- The evaluation covers 98 requirements reformulated into 146 requirements.

Health Monitor

- Responsible for detecting and addressing Errors.
- Includes 30 input requirements refined into 53 requirements.

Intra-Partition Communication

- Manages all communication between processes of a partition.
- Includes 68 input requirements refined into 93 requirements.

- ✓ They include requirements that differ in terms of content and structure.
- ✓ Their original requirements were reformulated using the Operational Behavior Template in our prior work (GT) *[1]*.
- ✓ We have already developed the domain models for these services in our previous work *[2]*.
- ✓ Both the specified requirements and domain models were validated by our industrial partner.

Experimental Configuration

RQ.1: **SOTA LLMs and prompting strategies for template-based requirement specification**

- ***Selected LLMs***: GPT, Llama, and Mistral.
- ***Prompting techniques***: zero-shot, few-shot, and Retrieval-Augmented Generation (RAG).

RQ.2: **Domain models as RAG context for template-based requirement specification**

- ***Domain model-enhanced RAG***: The LLM is provided with snippets from the domain model.
- ***Few-shot with RAG***: Comparing Few-shot & RAG with Few-shot & DM-enhanced RAG.

The evaluation covers a total of **138 experiments**.

Prompt

Persona & Requirements to Analyze

Instructions

Output Format

Template Definition

Template

You are a requirement engineer expert in the ARINC-653 certification. You will be provided with a functional requirement which was extracted from the Health Monitor section of the ARINC-653 specification.

Given the following instructions, template definitions, and domain model information, your task is to analyze the requirement, divide it if necessary, and reformulate it according to the template.

 # Requirement to Analyze:

#Instructions:
- If the requirement includes multiple subjects, action verbs, or objects within the action section, divide it into multiple requirements that each covers one subject, one action verb, and one object. You must generate all the requirements.
- If an element of the template is missing from the requirement in order to match the template, identify the correct information from the domain model.
- If a missing element cannot be identified from the domain model nor from the domain context, leave it empty and explain why in the rationale.
- Make the requirement verifiable when possible. If a condition or an action can be specified by a parameter and its value, replace it by the parameter attribute to make the requirement more verifiable.
- You must not add a condition if it is not specified in the input requirement or if it does not provide new information.
- If a requirement is already atomic and matches the template, do not change the wording unless it contradicts the Domain Model.
- The output requirement(s) should be written exactly according to the template. Do not add or remove any element from the template.
- The output requirement(s) should be in active voice.
- The output requirement(s) should not describe a different action or be semantically different from the input requirement.
- Provide a confidence score of the reformulated requirement(s), from a score of 0 to 3, with 0 being not confident at all and 3 being the most confident.
- Provide a rationale explaining all the decisions made during the reformulation. Provide only one paragraph for the rationale.

#Output Format
The output MUST be structured using these exact labels, each on a new line:
Reformulated Requirement(s): [Requirement(s)]
Rationale: [EXPLANATION]
Confidence Score: [SCORE]

 # Definition of the template:
- The template is used to specify functional requirements, describing an action that the system or its components perform under specific conditions. It includes two blocks: one block for conditions and another block for an action. The template allows zero or multiple conditions but only one action within the requirement.
- The template supports five types of conditions: (1) continuous action-driven conditions: describe a continuous action that triggers another action, (2) Discontinuous action-Driven conditions: describe a discontinuous action that can trigger another action, (3) Event-driven conditions: describe a constraint that is triggered by an event occurring at a specific moment of time, (4) Value-related conditions: specify whether a parameter is equal, inferior, or superior to a certain value, and (5) Range-related conditions: specify whether a parameter is within or out of a certain range.
- The template supports two types of actions: (1) operational actions specify an action that a subject shall perform on an object, and (2) value-related actions specify that a subject shall set a parameter to a certain value.
- The structure of the template is described in the textual template below. The templates for all types of conditions and actions are described in the predefined templates below.

 # Textual Template:
- The reformulated requirement MUST be structured using this exact template, with each line of the template placed on a new line:
[when / while / if
<condition>
[AND <condition>]*
Then]
<operational behavior action>

 # Predefined Templates:
<Continuous action-driven condition> : While <subject> [<detail>] <action verb(-ing)> [<object>] [<detail>]*
<Discontinuous action-driven condition> : When <object> [<detail>] <action verb(-ed)s> [[<adverb>] <object> [<detail>]]*
<Event-driven condition> : If <subject> [<detail>] <action verb> [[<adverb>] <object> [<detail>]]*
<Value-related condition> : If <parameter name> [of <entity> [<detail>]*] is equal/inferior/superior to <value>
<Range-related condition> : If <parameter name> [of <entity> [<detail>]*] is within / out of <value range>
<Operational action> : <subject> [<detail>]* shall <action verb> [[<adverb>] <object> [<detail>]]*
<Value-related action> : <subject> [<detail>]* shall set <parameter name> [of <entity> [<detail>]*] to <value>

<Subject> / **<Object>** : <range> <entity>
<Detail> : [<preposition>] [<range>] <entity>
<adverb> : for / as / from / to / by / upon / in
<value> : <parameter name> [of <entity> [<detail>]*] / <concrete value> / valid value / invalid value / infinite value
<range> : the / an / a / each / only one / all / one or more / at least one / other / another
<Preposition> : with / using / within / for / of / at / before / after / without / outside / in / on / via / until / to / other than / between / and
<Value range> : [<minimum value> , <maximum value>]

 # Notations clarification:
- < > denotes the variable parts that should be filled.
- [ ] denotes an optional block within the template.
- * denotes the possibility of having zero or many instances of a block.
- "/" denotes the options that can be used to fill the variable parts.

Metrics

- Evaluating template—based specification requires multiple dimensions: Logical decomposition, structural reasoning, and semantic grounding.

Logical Decomposition

Evaluates whether the requirement was correctly decomposed into atomic requirements.

- *Atomicity Precision (AP)*

$$AP = \frac{\text{Number of correctly generated atomic requirements}}{\text{Number of generated requirements}}$$

- *Atomicity Recall (AR)*

$$AR = \frac{\text{Number of correctly generated atomic requirements}}{\text{Number of expected atomic requirements}}$$

Structural reasoning

Evaluates whether the LLM selected the correct templates for conditions and actions.

- *Template Accuracy (TA)*

$$TA = \frac{\text{Number of correctly selected templates}}{\text{Number of templates in correctly decomposed requirements}}$$

Semantic Grounding

Evaluates the correctness of the information used to fill the placeholder slots within the condition and action templates.

- *Slot Filling Precision (SFP)*

$$SFP = \frac{\text{Number of correctly filled slots}}{\text{Number of slots in correctly selected templates}}$$

- *Overall Correctness :* $OC = \alpha \cdot AP + \beta \cdot TA + \gamma \cdot SFP$ (α=0.2 β=0.4 γ=0.4)

Experimental Results

RQ.1: LLMs and Prompting Strategies

Zero-shot

LLM	HM Dataset					IPC Dataset				
	AP	*AR*	*TA*	*SFP*	*OC*	*AP*	*AR*	*TA*	*SFP*	*OC*
GPT-5.4	0.76	0.64	0.80	0.56	0.70	0.80	0.82	0.75	0.44	0.64
Llama	0.45	0.42	0.61	0.37	0.48	0.42	0.51	0.56	0.38	0.46
Mistral	0.57	0.38	0.63	0.40	0.53	0.66	0.57	0.67	0.30	0.52

Few-shot

LLM	No. Examples	HM Dataset					IPC Dataset				
		AP	*AR*	*TA*	*SFP*	*OC*	*AP*	*AR*	*TA*	*SFP*	*OC*
GPT-5.4	**2 Examples**	0.66	0.58	0.79	0.67	0.72	**0.83**	0.84	0.82	0.63	0.75
	4 Examples	0.64	0.57	0.79	0.63	0.70	**0.83**	0.84	0.86	0.72	0.80
	6 Examples	0.67	0.60	0.77	0.69	0.72	0.80	0.81	0.84	0.73	0.79
	8 Examples	0.71	0.64	0.85	0.67	0.75	0.82	0.83	0.89	0.78	0.83
	10 Examples	0.62	0.58	0.75	**0.77**	0.73	0.82	**0.85**	0.87	0.74	0.81
Llama	**2 Examples**	0.62	0.58	0.73	0.31	0.54	0.30	0.35	0.71	0.54	0.56
	4 Examples	0.47	0.40	0.67	0.41	0.53	0.33	0.35	0.76	0.57	0.60
	6 Examples	0.31	0.30	0.65	0.53	0.53	0.40	0.46	0.69	0.57	0.58
	8 Examples	0.29	0.28	0.56	0.57	0.51	0.31	0.37	0.71	0.55	0.57
	10 Examples	0.35	0.32	0.58	0.18	0.37	0.32	0.35	0.67	0.60	0.57
Mistral	**2 Examples**	0.47	0.38	0.72	0.45	0.56	0.37	0.34	0.76	0.56	0.60
	4 Examples	0.31	0.23	0.75	0.51	0.57	0.26	0.24	0.69	0.34	0.46
	6 Examples	0.42	0.28	0.75	0.39	0.54	0.27	0.24	0.79	0.45	0.55
	8 Examples	0.41	0.30	0.76	0.55	0.61	0.35	0.32	0.73	0.54	0.58
	10 Examples	0.49	0.34	0.69	0.55	0.59	0.23	0.19	0.76	0.58	0.58

RAG

LLM	Chunk Size	No. Chunks	HM Dataset					IPC Dataset				
			AP	*AR*	*TA*	*SFP*	*OC*	*AP*	*AR*	*TA*	*SFP*	*OC*
GPT-5.4	500c	**2 Chunks**	0.65	0.53	0.80	0.56	0.67	0.72	0.75	0.68	0.53	0.63
		4 Chunks	0.66	0.51	0.80	0.62	0.70	0.78	0.78	0.73	0.47	0.64
		6 Chunks	0.80	0.66	0.64	0.60	0.66	0.78	0.78	0.71	0.49	0.64
		8 Chunks	0.84	0.58	0.70	0.67	**0.72**	0.80	0.77	**0.75**	0.48	0.65
		10 Chunks	0.87	0.77	0.82	0.52	0.71	0.79	0.80	0.72	0.53	0.66
	1000c	**2 Chunks**	0.77	0.62	0.72	0.58	0.67	0.85	**0.83**	0.69	0.47	0.63
		4 Chunks	0.63	0.51	0.80	0.48	0.64	0.82	0.81	0.68	0.44	0.61
		6 Chunks	0.66	0.51	0.72	0.52	0.63	0.76	0.76	0.70	0.48	0.62
		8 Chunks	0.62	0.47	0.75	**0.72**	0.71	**0.86**	0.81	0.71	0.51	**0.66**
		10 Chunks	0.79	0.62	0.78	0.56	0.69	0.82	0.81	0.66	**0.58**	**0.66**
Llama	500c	**2 Chunks**	0.54	0.40	0.56	0.48	0.52	0.33	0.40	0.70	0.24	0.44
		4 Chunks	0.39	0.34	0.68	0.42	0.52	0.51	0.56	0.65	0.27	0.47
		6 Chunks	0.49	0.40	0.57	0.34	0.46	0.50	0.56	0.66	0.29	0.48
		8 Chunks	0.34	0.28	0.67	0.46	0.52	0.47	0.55	0.62	0.38	0.49
		10 Chunks	0.33	0.30	0.48	0.39	0.41	0.41	0.48	0.62	0.33	0.46
	1000c	**2 Chunks**	0.42	0.34	0.56	0.32	0.44	0.48	0.46	0.57	0.27	0.43
		4 Chunks	0.43	0.40	0.54	0.34	0.44	0.49	0.53	0.66	0.25	0.46
		6 Chunks	0.37	0.30	0.48	0.35	0.41	0.45	0.48	0.64	0.26	0.45
		8 Chunks	0.29	0.28	0.55	0.30	0.40	0.41	0.45	0.56	0.16	0.37
		10 Chunks	0.40	0.32	0.65	0.29	0.46	0.38	0.42	0.64	0.23	0.42
Mistral	500c	**2 Chunks**	0.38	0.30	0.66	0.25	0.44	0.54	0.56	0.69	0.29	0.50
		4 Chunks	0.44	0.34	0.68	0.27	0.47	0.47	0.41	0.67	0.28	0.47
		6 Chunks	0.63	0.49	0.66	0.34	0.53	0.52	0.51	0.75	0.30	0.52
		8 Chunks	0.52	0.42	0.72	0.31	0.52	0.55	0.53	0.65	0.25	0.47
		10 Chunks	0.50	0.40	0.65	0.28	0.47	0.50	0.43	0.73	0.22	0.48
	1000c	**2 Chunks**	0.45	0.32	0.72	0.24	0.47	0.52	0.54	0.73	0.27	0.50
		4 Chunks	0.46	0.36	0.64	0.29	0.46	0.54	0.48	0.72	0.20	0.48
		6 Chunks	0.34	0.28	0.62	0.30	0.44	0.44	0.45	0.63	0.21	0.42
		8 Chunks	0.37	0.36	0.61	0.22	0.41	0.43	0.43	0.63	0.24	0.43
		10 Chunks	0.37	0.34	0.69	0.23	0.44	0.36	0.31	0.67	0.19	0.42

- ✓ GPT-5.4 requires a relatively large number of examples to effectively guide the task in few-shot prompting.
- ✓ GPT-5.4 requires a relatively large number of 500-character chunks to provide sufficient context through RAG.
- ✓ Few-shot prompting provides a more significant improvement over zero-shot and traditional RAG, highlighting the importance of illustrative examples for this task.

RQ.2: Standard RAG vs DM-Enhanced RAG

Standard RAG

LLM	Chunk Size	No. Chunks	HM Dataset					IPC Dataset				
			AP	*AR*	*TA*	*SFP*	*OC*	*AP*	*AR*	*TA*	*SFP*	*OC*
GPT-5.4	500c	2 Chunks	0.65	0.53	0.80	0.56	0.67	0.72	0.75	0.68	0.53	0.63
		4 Chunks	0.66	0.51	0.80	0.62	0.70	0.78	0.78	0.73	0.47	0.64
		6 Chunks	0.80	0.66	0.64	0.60	0.66	0.78	0.78	0.71	0.49	0.64
		8 Chunks	0.84	0.58	0.70	0.67	**0.72**	0.80	0.77	**0.75**	0.48	0.65
		10 Chunks	**0.87**	**0.77**	**0.82**	0.52	0.71	0.79	0.80	0.72	0.53	**0.66**
	1000c	2 Chunks	0.77	0.62	0.72	0.58	0.67	0.85	**0.83**	0.69	0.47	0.63
		4 Chunks	0.63	0.51	0.80	0.48	0.64	0.82	0.81	0.68	0.44	0.61
		6 Chunks	0.66	0.51	0.72	0.52	0.63	0.76	0.76	0.70	0.48	0.62
		8 Chunks	0.62	0.47	0.75	**0.72**	0.71	0.86	0.81	0.71	0.51	**0.66**
		10 Chunks	0.79	0.62	0.78	0.56	0.69	0.82	0.81	0.66	**0.58**	**0.66**
Llama	500c	2 Chunks	0.54	0.40	0.56	0.48	0.52	0.33	0.40	0.70	0.24	0.44
		4 Chunks	0.39	0.34	0.68	0.42	0.52	0.51	0.56	0.65	0.27	0.47
		6 Chunks	0.49	0.40	0.57	0.34	0.46	0.50	0.56	0.66	0.29	0.48
		8 Chunks	0.34	0.28	0.67	0.46	0.52	0.47	0.55	0.62	0.38	0.49
		10 Chunks	0.33	0.30	0.48	0.39	0.41	0.41	0.48	0.62	0.33	0.46
	1000c	2 Chunks	0.42	0.34	0.56	0.32	0.44	0.48	0.46	0.57	0.27	0.43
		4 Chunks	0.43	0.40	0.54	0.34	0.44	0.49	0.53	0.66	0.25	0.46
		6 Chunks	0.37	0.30	0.48	0.35	0.41	0.45	0.48	0.64	0.26	0.45
		8 Chunks	0.29	0.28	0.55	0.30	0.40	0.41	0.45	0.56	0.16	0.37
		10 Chunks	0.40	0.32	0.65	0.29	0.46	0.38	0.42	0.64	0.23	0.42
Mistral	500c	2 Chunks	0.38	0.30	0.66	0.25	0.44	0.54	0.56	0.69	0.29	0.50
		4 Chunks	0.44	0.34	0.68	0.27	0.47	0.47	0.41	0.67	0.28	0.47
		6 Chunks	0.63	0.49	0.66	0.34	0.53	0.52	0.51	0.75	0.30	0.52
		8 Chunks	0.52	0.42	0.72	0.31	0.52	0.55	0.53	0.65	0.25	0.47
		10 Chunks	0.50	0.40	0.65	0.28	0.47	0.50	0.43	0.73	0.22	0.48
	1000c	2 Chunks	0.45	0.32	0.72	0.24	0.47	0.52	0.54	0.73	0.27	0.50
		4 Chunks	0.46	0.36	0.64	0.29	0.46	0.54	0.48	0.72	0.20	0.48
		6 Chunks	0.34	0.28	0.62	0.30	0.44	0.44	0.45	0.63	0.21	0.42
		8 Chunks	0.37	0.36	0.61	0.22	0.41	0.43	0.43	0.63	0.24	0.43
		10 Chunks	0.37	0.34	0.69	0.23	0.44	0.36	0.31	0.67	0.19	0.42

DM-Enhanced RAG

LLM	No. Chunks	HM Dataset					IPC Dataset				
		AP	*AR*	*TA*	*SFP*	*OC*	*AP*	*AR*	*TA*	*SFP*	*OC*
GPT-5.4	5 Chunks	0.60	0.51	0.78	0.53	0.64	0.82	0.83	**0.75**	0.48	0.66
	10 Chunks	0.69	0.55	0.73	**0.62**	0.68	0.76	0.77	0.65	0.49	0.61
	15 Chunks	0.66	0.55	0.78	0.59	0.68	0.80	0.80	0.72	0.46	0.63
	20 Chunks	0.68	0.57	0.71	0.59	0.66	0.79	0.82	0.73	0.50	0.65
	Entire Model	0.92	0.85	0.85	0.54	0.74	0.84	0.83	0.70	0.55	0.67
Llama	5 Chunks	0.64	0.53	0.59	0.34	0.50	0.50	0.56	0.58	0.22	0.42
	10 Chunks	0.43	0.34	0.70	0.36	0.51	0.55	0.62	0.55	0.25	0.43
	15 Chunks	0.45	0.38	0.62	0.40	0.50	0.53	0.56	0.60	0.35	0.49
	20 Chunks	0.57	0.45	0.57	0.32	0.47	0.41	0.47	0.47	0.21	0.35
	Entire Model	0.29	0.28	0.62	0.33	0.44	0.36	0.39	0.58	0.17	0.37
Mistral	5 Chunks	0.59	0.42	0.71	0.41	0.57	0.61	0.59	0.62	0.23	0.46
	10 Chunks	0.36	0.25	0.68	0.26	0.45	0.63	0.68	0.64	0.23	0.47
	15 Chunks	0.54	0.40	0.67	0.25	0.48	0.49	0.55	0.61	0.20	0.42
	20 Chunks	0.39	0.30	0.53	0.11	0.33	0.59	0.66	0.59	0.21	0.44
	Entire Model	0.00	0.00	0.00	0.00	0.00	0.06	0.03	0.33	0.00	0.14

Execution Time

LLM	HM Dataset		IPC Dataset	
	RAG	DM-RAG	RAG	DM-RAG
GPT-5.4	195 s	177 s	470 s	441 s
Llama	328 s	289 s	545 s	393 s
Mistral	257 s	233 s	355 s	352 s

- ✓ While DM-enhanced RAG reduces execution time compared to traditional RAG, the performance gains remain marginal.
- ✓ Few-shot prompting remains the most effective strategy for template-based requirements specification.

RQ.2: Standard RAG vs DM-Enhanced RAG

Combining RAG with Few-shot

RAG	LLM	HM Dataset					IPC Dataset				
		AP	*AR*	*TA*	*SFP*	*OC*	*AP*	*AR*	*TA*	*SFP*	*OC*
Standard	GPT-5.4	**0.83**	**0.85**	0.72	**0.60**	0.69	0.81	0.83	**0.90**	0.70	0.80
	Llama	0.21	0.17	0.64	0.24	0.39	0.26	0.31	0.79	0.60	0.61
	Mistral	0.47	0.34	0.78	0.33	0.54	0.26	0.24	0.88	0.36	0.55
DM	GPT-5.4	0.80	**0.85**	**0.85**	0.52	**0.71**	**0.84**	**0.86**	0.89	**0.72**	**0.81**
	Llama	0.42	0.40	0.74	0.53	0.59	0.32	0.32	0.74	0.56	0.58
	Mistral	0.49	0.34	0.73	0.51	0.59	0.24	0.22	0.79	0.50	0.56

Few-shot

LLM	No. Examples	HM Dataset					IPC Dataset				
		AP	*AR*	*TA*	*SFP*	*OC*	*AP*	*AR*	*TA*	*SFP*	*OC*
GPT-5.4	2 Examples	0.66	0.58	0.79	0.67	0.72	**0.83**	0.84	0.82	0.63	0.75
	4 Examples	0.64	0.57	0.79	0.63	0.70	**0.83**	0.84	0.86	0.72	0.80
	6 Examples	0.67	0.60	0.77	0.69	0.72	0.80	0.81	0.84	0.73	0.79
	8 Examples	**0.71**	**0.64**	**0.85**	0.67	**0.75**	0.82	0.83	**0.89**	**0.78**	**0.83**
	10 Examples	0.62	0.58	0.75	**0.77**	0.73	0.82	**0.85**	0.87	0.74	0.81

- ✓ Combining few-shot with RAG does not provide any benefits.
- ✓ Simple few-shot suffices, yielding the highest results across all conducted experiments.

Threats to Validity

Internal Validity:

The GT and domain models were manually constructed by one of the authors.

- ✓ Both the GT and domain models were validated by an RE expert and our industrial partner.

The evaluation of metrics was performed by the authors.

- ✓ Metrics evaluation was conducted through a script that strictly adheres to the guidelines.

Variability of SCS documents.

- ✓ SCS generally rely on a more constrained vocabulary.

External Validity:

The evaluation is limited to a single template.

- ✓ The template is general and applicable across domains, as demonstrated in our previous work *[2]*.

The evaluation is limited to the ARINC-653 standard.

- ✓ The datasets include requirements with varying structures and contents.

Conclusion

Conclusion

Overview

- The paper presents an empirical evaluation of LLM-enabled template-based requirements specification.
- We compared various prompting strategies and context-enrichment techniques, including zero-shot, few-shot, traditional RAG, and DM-enhanced RAG.
- The evaluation covered 138 requirements from two datasets from the ARINC-653 standard, assessing logical decomposition, structural reasoning, and semantic grounding.

Results

- GPT-5.4 outperformed Mistral and Llama, with Mistral showing better overall performance than Llama.
- Leveraging domain models yielded comparable overall correctness to standard RAG, suggesting that the overhead of creating and maintaining domain models may not be justified for the task.
- Few-shot prompting achieved the highest performance.
- Combining few-shot with RAG proved less effective than few-shot alone, demonstrating that illustrative examples are sufficient.

Future Work

- Compare the cost of LLMs against manual effort to assess their value in industrial settings.
- Explore the use of a multi-agent architecture to support the various specification steps.
- Extend the study to include other safety-critical domains.

Bibliography

[1] Darif, I., El Boussaidi, G. & Kpodjedo, S. Requirements specification using templates: a model-driven approach. Softw Syst Model 24, 1897–1934 (2025). https://doi.org/10.1007/s10270-025-01265-6.

[2] SAE, “ARINC Specification653P1-4. Avionics application software standard interface,” p. 285, 2015.



THANK YOU FOR YOUR ATTENTION!