Benchmarking Generative AI for Requirements Management

Ruslan Bernijazov
& Nikhil Jha

About us

Speakers of the session

RUSLAN BERNIJAZOV

MD & Co-founder, AI Marketplace GmbH
10+ Years Experience in Systems Engineering
Fraunhofer IEM | Heinz Nixdorf Institute

NIKHIL JHA

AI Technology Generalist, AI Marketplace GmbH
Lead, Google Developers Group Paderborn
Fraunhofer IEM | Wipro

About us

Engineering slows down as complexity grows

The daily routine of an engineer:

The daily problems of an engineer:

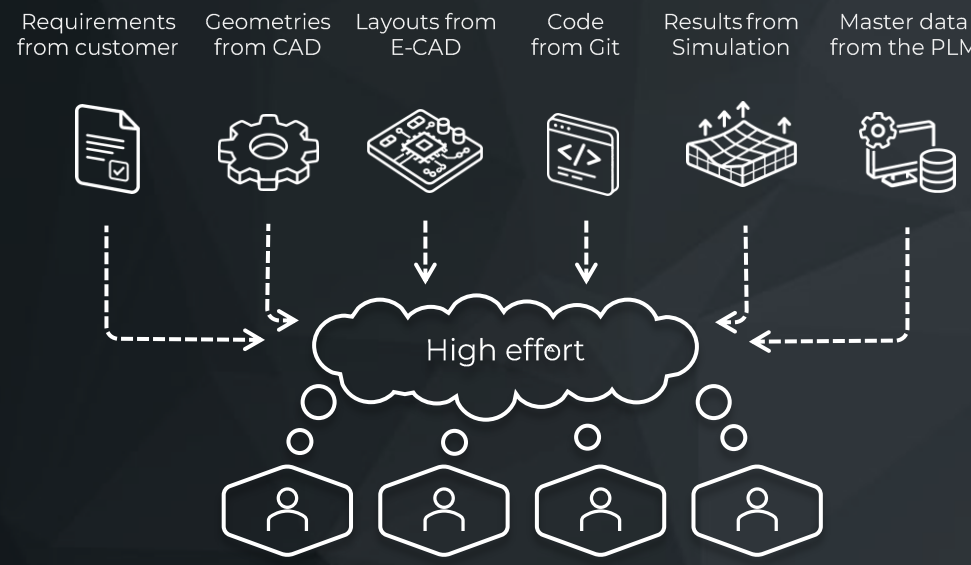
1. Change impact analysis across requirements & architecture takes weeks
2. Engineering data is fragmented across tools
3. Critical artifacts (requirements, architectures, simulation models) are not AI-ready

About us

We’re building the platform that operates agents in engineering

Marketplace

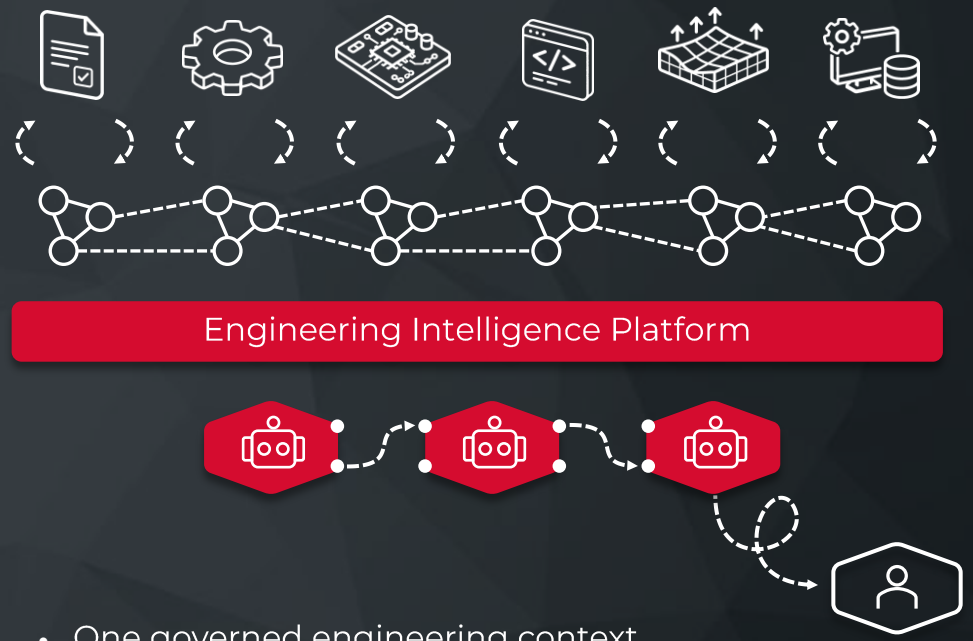
Current way



- Engineer searches ReqIF, SysML, PDFs
- Manual cross-tool reasoning
- High risk of missing impacts

Our vision

and why it’s better



- One governed engineering context
- Agent analyzes impact
- Human approves → traceable decision

Slide 4

Motivation

Challenges in Requirements Management

Marketplace

- Manual Process
- Error Prone
- Regulations
- Data Constraints
- Compliance

Motivation

Challenges in Requirements Management

Marketplace

Manual Process

Error Prone

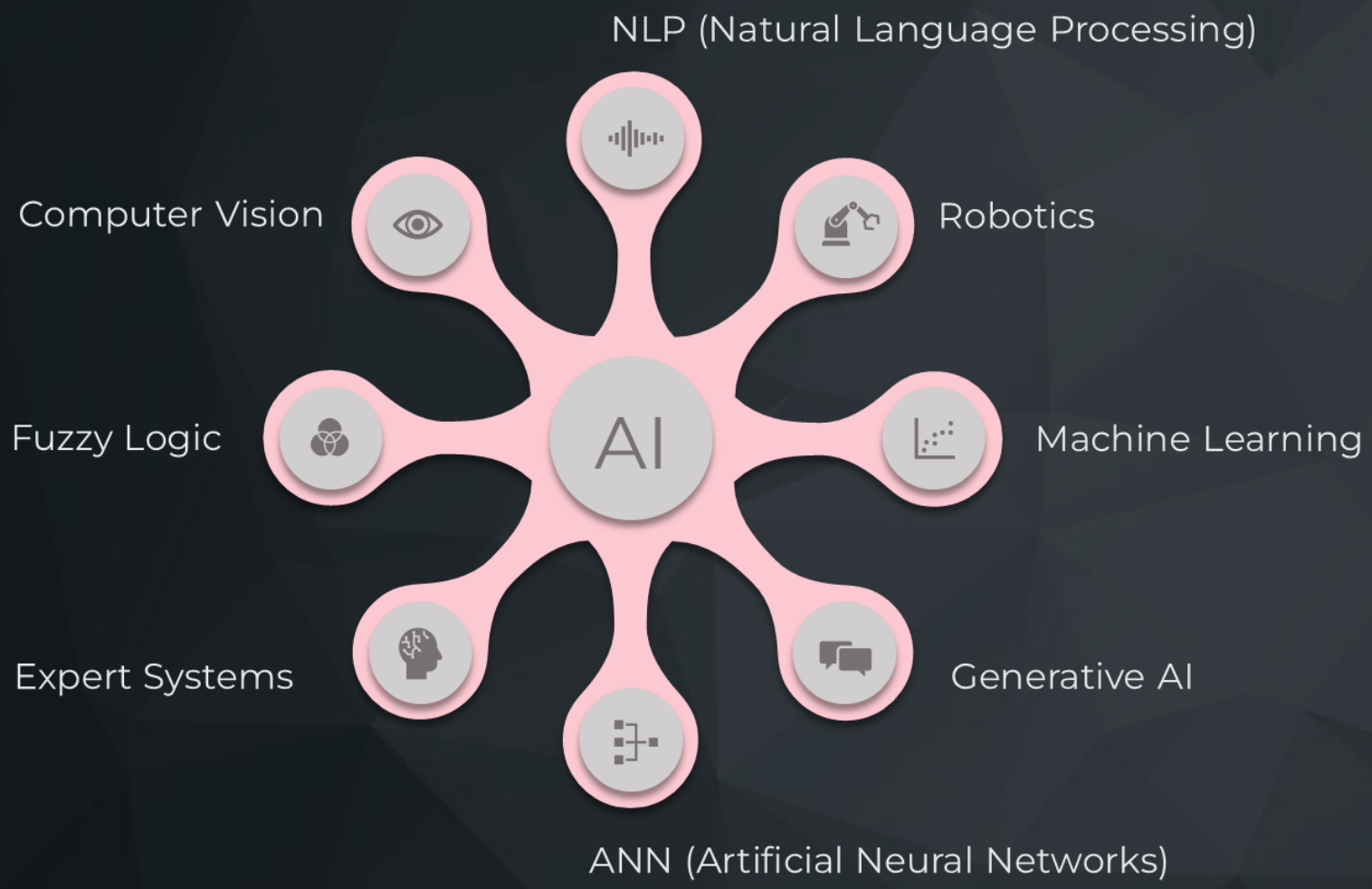
Regulations

Data Constraints

Compliance

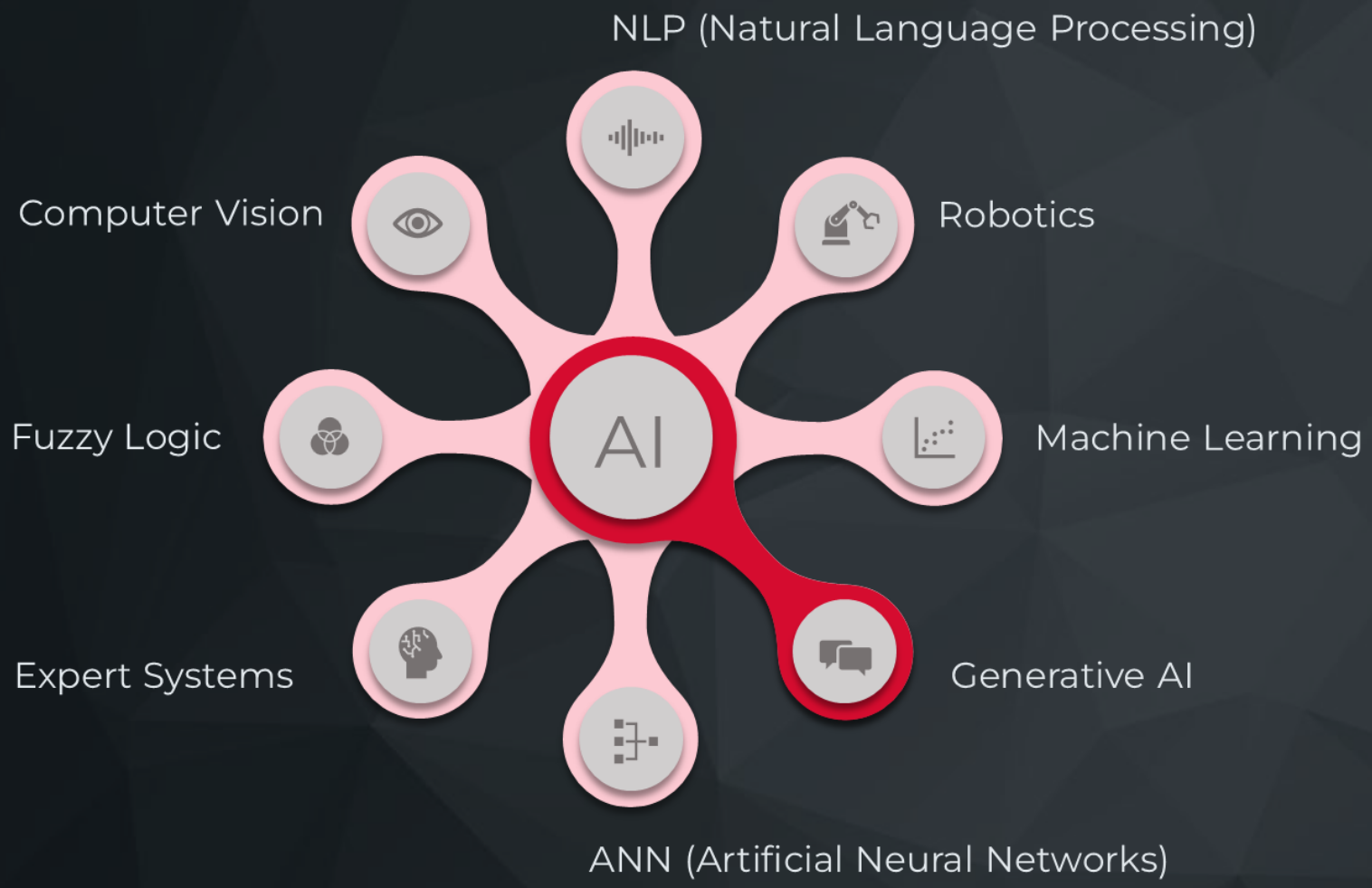
Artificial Intelligence

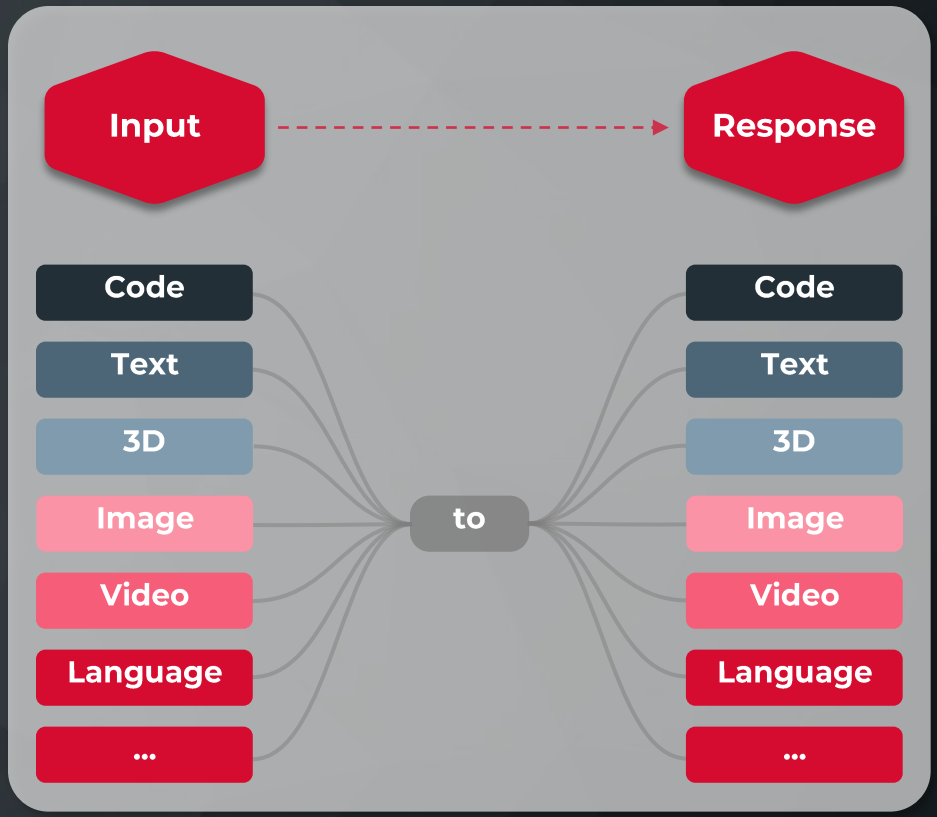
Fields of Artificial Intelligence (AI)



Motivation

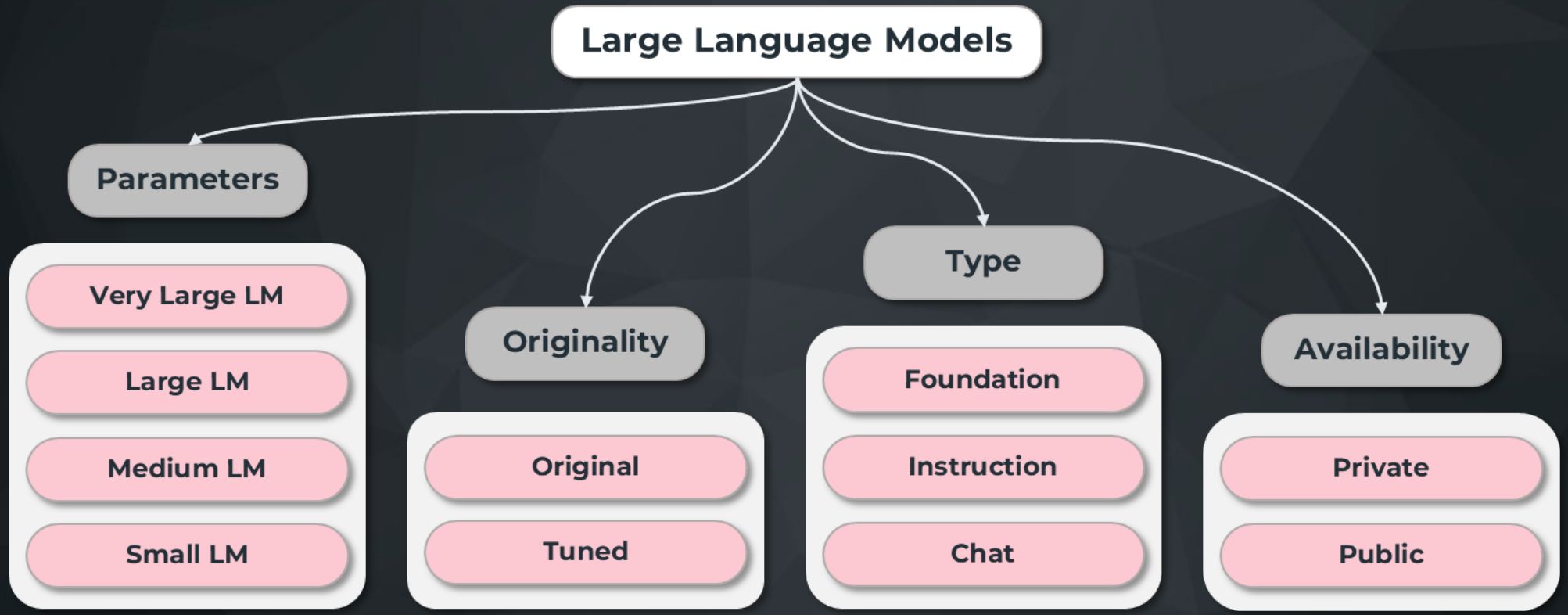
Generative AI





Motivation

Large Language Models



Source: Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., & Gao, J. (2024). Large language models: A survey. arXiv preprint arXiv:2402.06196.

Motivation

Prompting Techniques

Marketplace

- Zero-Shot Promting
- Few-Shot Promting
- Chain-of-thought
- Tree-of-thought
- Iterative refinement
- Feedback loop
- Prompt chaining
- Role-playing
- Maieutic Prompting
- Complexity based Prompting

https://www.servicenow.com/de/ai/what-is-prompt-engineering.html

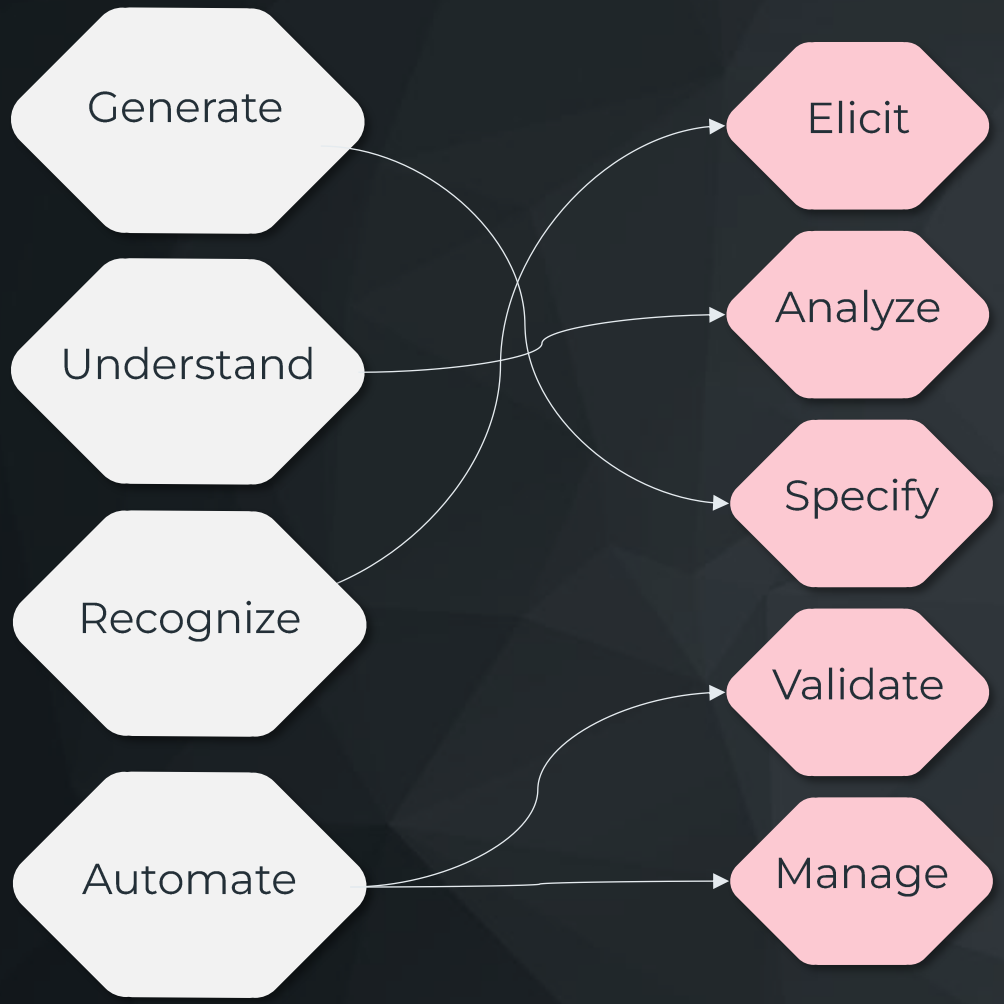
Motivation

Why Generative AI ?

- Generate
- Understand
- Recognize
- Automate

Motivation

Why Generative AI in RM?



Advancing Requirements Engineering Through Generative AI: Assessing the Role of LLMs

Chetan Arora, John Grundy, and Mohamed Abdelrazek

Abstract Requirements Engineering (RE) is a cr
ment including the elicitation, analysis, specific
requirements. Despite the importance of RE, it ren
the complexities of communication, uncertainty i
automation support. In recent years, large langu
significant promise in diverse domains, includi
code generation, and program understanding. Th
of LLMs in driving RE processes, aiming to img
of requirements-related tasks. We propose key di
research and development in using LLMs for R
requirements elicitation, analysis, specification, a
the results from a preliminary evaluation, in this c

Keywords Requirements engineering · Generati
(LLMs) · Natural language processing · Softwar

1 Introduction

Requirements Engineering (RE) is arguably the
development process, where the needs and const
analyzed, and documented to create a well-define
nizations and project teams often overlook or d
of RE and its impact on project success [24]. S
lack of effort and resources spent in RE includ

C. Arora (✉) · J. Grundy
Monash University, Melbourne, VIC, Australia
e-mail: chetan.arora@monash.edu; john.grundy@monash.

M. Abdelrazek
Deakin University, Geelong, VIC, Australia
e-mail: mohamed.abdelrazek@deakin.edu.au

A. Nguyen-Duc et al. (eds.), *Generative AI for Effective So*
https://doi.org/10.1007/978-3-031-55642-5_6

Requirements Engineering Using Generative AI: Prompts and Prompting Patterns

Krishna Ronanki, Beatriz Cabrero-Daniel, Jennifer Horkoff and Christian Berger

Abstract **[Context]** Companies are increasingly recognizing the i
automating Requirements Engineering (RE) tasks due to their resou
nature. The advent of GenAI has made these tasks more amenable t
thanks to its ability to understand and interpret context effectively
However, in the context of GenAI, prompt engineering is a critical facto
Despite this, we currently lack tools and methods to systematicall
determine the most effective prompt patterns to employ for a partic
[Method] Two tasks related to requirements, specifically requirement
and tracing, were automated using the GPT-3.5 turbo API. The
evaluation involved assessing various prompts created using 5 pro
and implemented programmatically to perform the selected RE ta
on metrics such as precision, recall, accuracy, and F-Score. **[Result**
evaluates the effectiveness of the 5 prompt patterns' ability to make C
perform the selected RE tasks and offers recommendations on which p
to use for a specific RE task. Additionally, it also provides an evaluatio
as a reference for researchers and practitioners who want to evalu
prompt patterns for different RE tasks.

Keywords Requirements engineering · Generative AI · Prompt patt
Prompt engineering · Large language models

1 Introduction

Researchers in Requirements Engineering (RE) have been exploring the use of machine learning (ML) and deep learning (DL) methods for various RE tasks, including requirements classification, prioritization, tracing, ambiguity detection, and modelling [1]. However, the majority of existing ML/DL approaches are based

K. Ronanki (✉) · B. Cabrero-Daniel · J. Horkoff · C. Berger
University of Gothenburg, Gothenburg, Sweden
e-mail: krishna.ronanki@gu.se

A. Nguyen-Duc et al. (eds.), *Generative AI for Effective Software Development*,
https://doi.org/10.1007/978-3-031-55642-5_5

Generative AI for Requirements Engineering: A Systematic Literature Review

Haowei Cheng[a], Jati H. Husen[a,b], Sien Reeve Peralta[a], Bowen Jiang[a], Nobukazu Yoshioka[a], Naoyasu Ubayashi[a], Hironori Washizaki[a]

[a]*Waseda University, Tokyo, 169-8050, Japan*
[b]*Telkom University, Bandung, 140257, Jawa Barat, Indonesia*

Abstract

Context: Generative AI (GenAI) has emerged as a transformative tool in software engineering, with requirements engineering (RE) actively exploring its potential to revolutionize processes and outcomes. The integration of GenAI into RE presents both promising opportunities and significant challenges that necessitate systematic analysis and evaluation.
Objective: This paper presents a comprehensive systematic literature review (SLR) analyzing state-of-the-art applications and innovative proposals leveraging GenAI in RE. It surveys studies focusing on the utilization of GenAI to enhance RE processes while identifying key challenges and opportunities in this rapidly evolving field.
Method: A rigorous SLR methodology was used to conduct an in-depth analysis of 27 carefully selected primary studies. The review examined research questions pertaining to the application of GenAI across various RE phases, the models and techniques used, and the challenges encountered in implementation and adoption.
Results: The most salient findings include i) a predominant focus on the early stages of RE, particularly the elicitation and analysis of requirements, indicating potential for expansion into later phases; ii) the dominance of large language models, especially the GPT series, highlighting the need for diverse AI approaches; and iii) persistent challenges in domain-specific applications and the interpretability of AI-generated outputs, underscoring areas requiring further research and development.
Conclusions: The results highlight the critical need for comprehensive evaluation frameworks, improved human–AI collaboration models, and thorough consideration of ethical implications in GenAI-assisted RE. Future research should prioritize extending GenAI applications across the entire RE lifecycle, enhancing domain-specific capabilities, and developing strategies for responsible AI integration in RE practices.

Keywords: Generative AI, Requirements Engineering, Systematic Literature

Preprint submitted to Information and Software Technology *September 12, 2024*

Motivation

Limitations of Generative AI

Hallucination

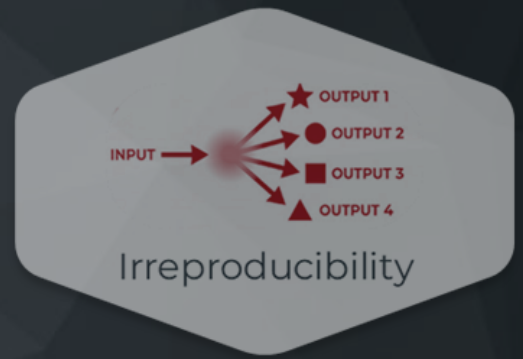
Coherence Loss

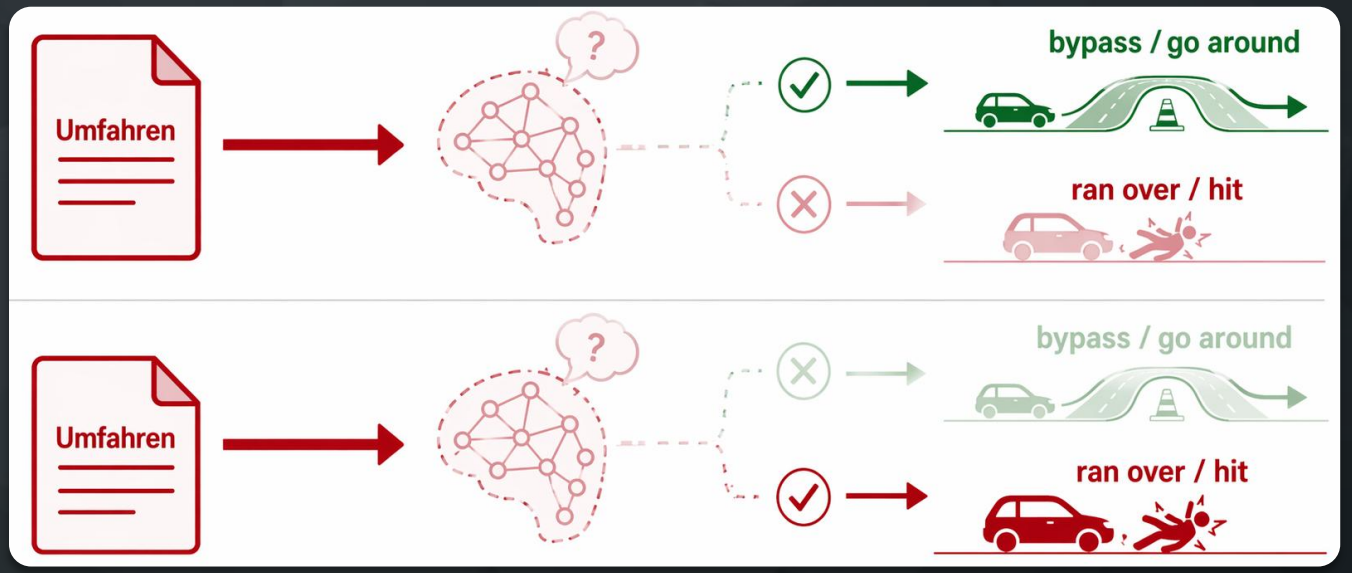
Black Box

Irreproducibility

Limitations of Generative AI

Marketplace






THE GREAT
GenAI RACE
M
L
G
D
C
BENCHMARKS

Benchmarks

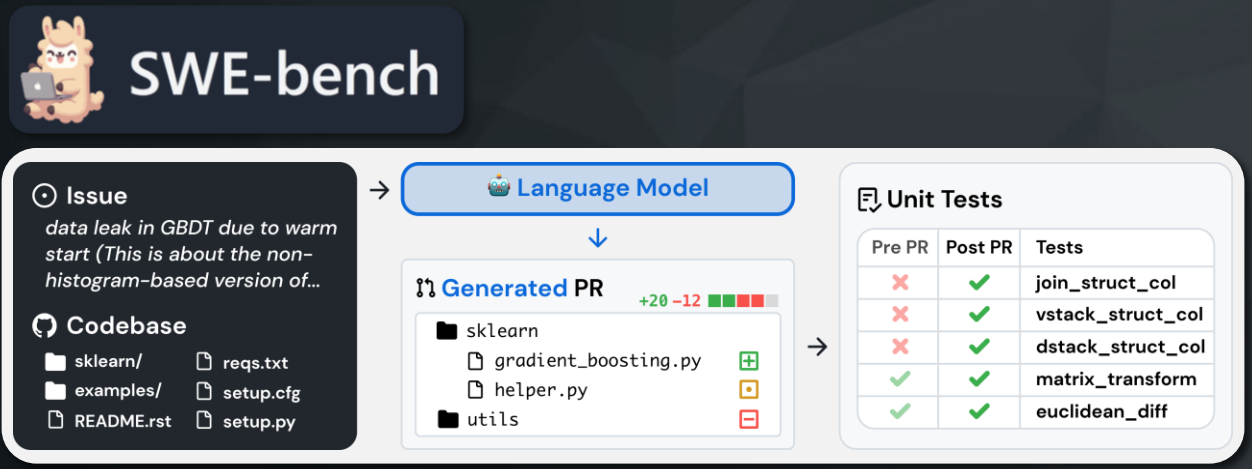
What is Benchmark ?

Marketplace

A scientific way for **evaluating and comparing** system performance **under standardized conditions.**

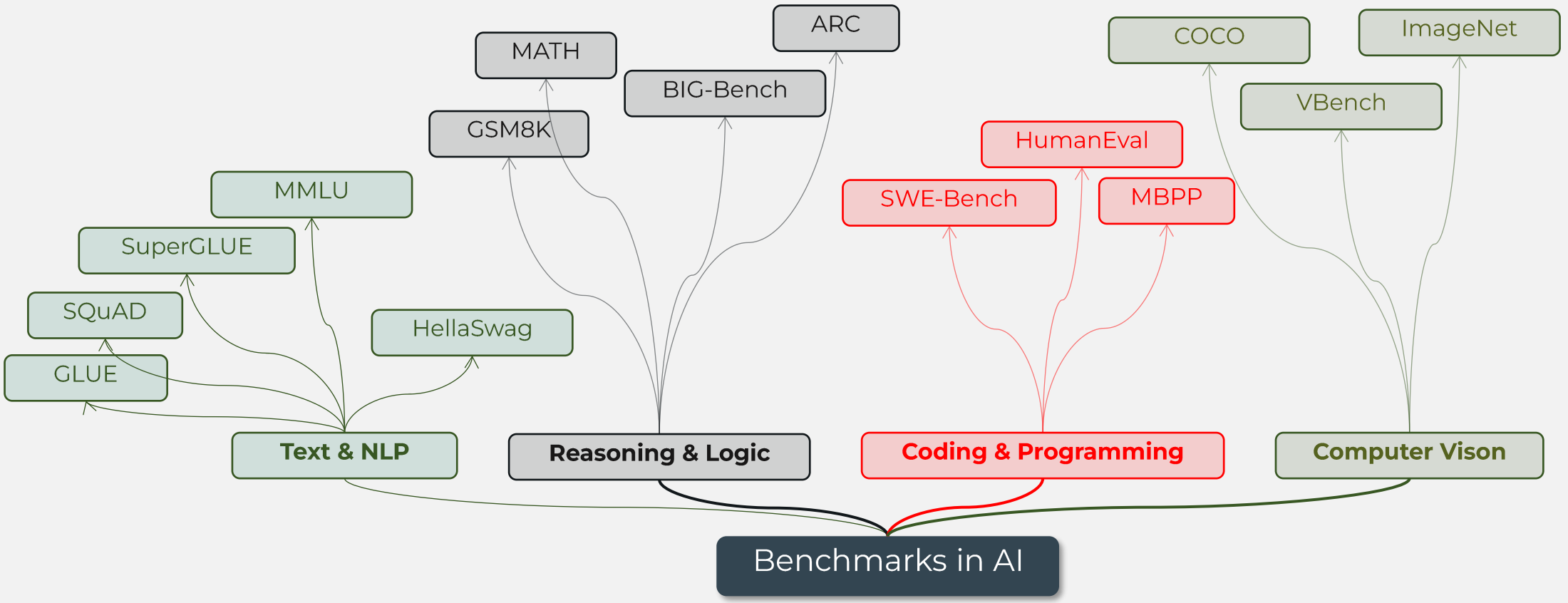
Workload | Methodology | Metrics

SWE-bench

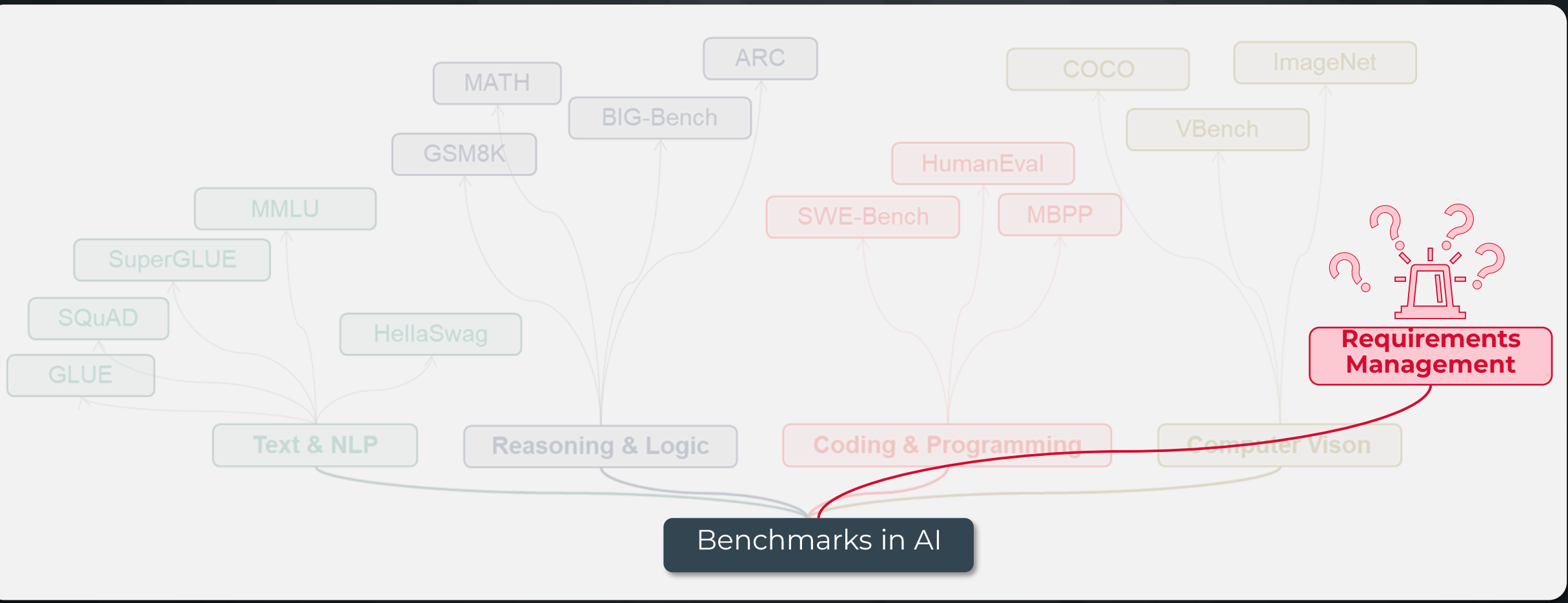


Model	% Resolved	Avg. $	Trajs	Org	Date	Agent
Claude 4.5 Opus (high reasoning)	76.80	$0.75			2026-02-17	2.0.0
Gemini 3 Flash (high reasoning)	75.80	$0.36			2026-02-17	2.0.0
MiniMax M2.5 (high reasoning)	75.80	$0.07			2026-02-17	2.0.0
Claude Opus 4.6	75.60	$0.55			2026-02-17	2.0.0
GPT-5-2 Codex	72.80	$0.45			2026-02-19	2.0.0
GLM-5 (high reasoning)	72.80	$0.53			2026-02-17	2.0.0
GPT-5-2 (high reasoning)	72.80	$0.47			2026-02-17	2.0.0
GPT 5.2 Codex	72.80	$0.45			2026-02-19	2.0.0
Claude 4.5 Sonnet (high reasoning)	71.40	$0.66			2026-02-17	2.0.0
Kimi K2.5 (high reasoning)	70.80	$0.15			2026-02-17	2.0.0
DeepSeek V3.2 (high reasoning)	70.00	$0.45			2026-02-17	2.0.0
Gemini 3 Pro	69.60	$0.96			2026-02-26	2.0.0
Claude 4.5 Haiku (high reasoning)	66.60	$0.33			2026-02-17	2.0.0
GPT-5 Mini	56.20	$0.05			2026-02-17	2.0.0

Benchmark for AI in Requirements Management ?



Benchmark for AI in Requirements Management ?



Our Solution : RM Bench

Marketplace

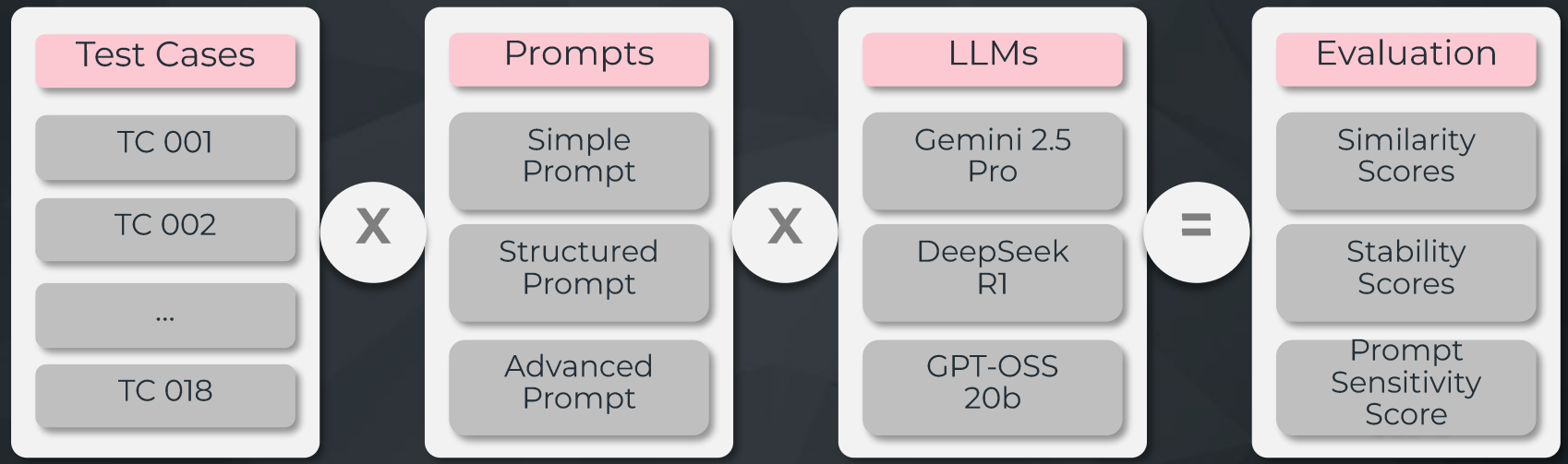
Change Management

18 Test Cases

3 Prompt Types

13 GenAI Models

LLM as a Judge



RM Bench

Test Case Example

Marketplace

Initial Input

The drone shall achieve a mean time between critical failures (MTBCF) of **at least 200 flight-hours**.

Change Logs

UAVs shall demonstrate a mean time between critical failures (MTBCF) of **no less than 300 flight-hours**, reflecting improved hardware reliability and customer expectations for longer uptime.

Prompt

Update the list of requirements by changing, adding, or removing them exactly as told, and then give back the full updated list in the same format.

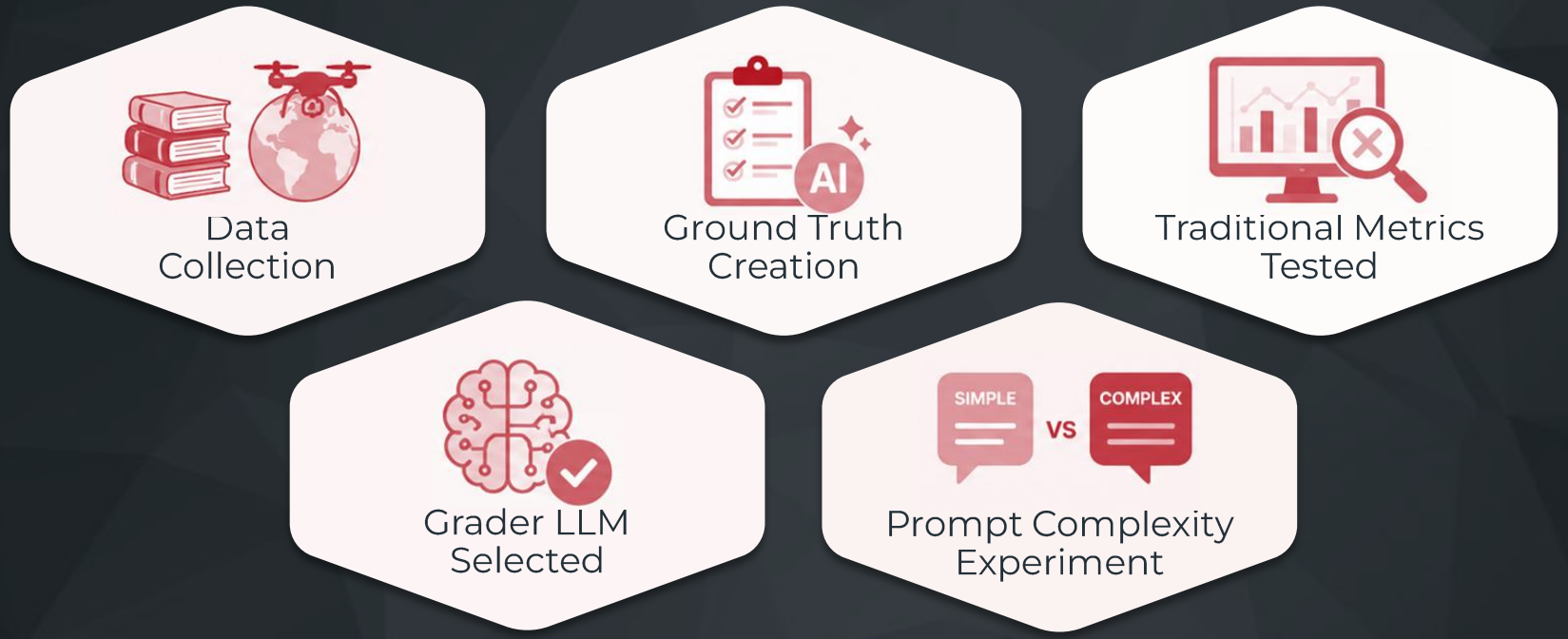
Golden Output

The drone shall achieve a mean time between critical failures (MTBCF) of **at least 300 flight-hours.**

RM Bench

Experiment Setup

Marketplace

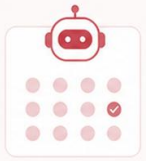


RM Bench

How it works ?

Marketplace

1
Test-Case Input Artefacts

2

LLM Candidate Selection

3


Prompt Variation Loop

4


Model Outputs Saved

5

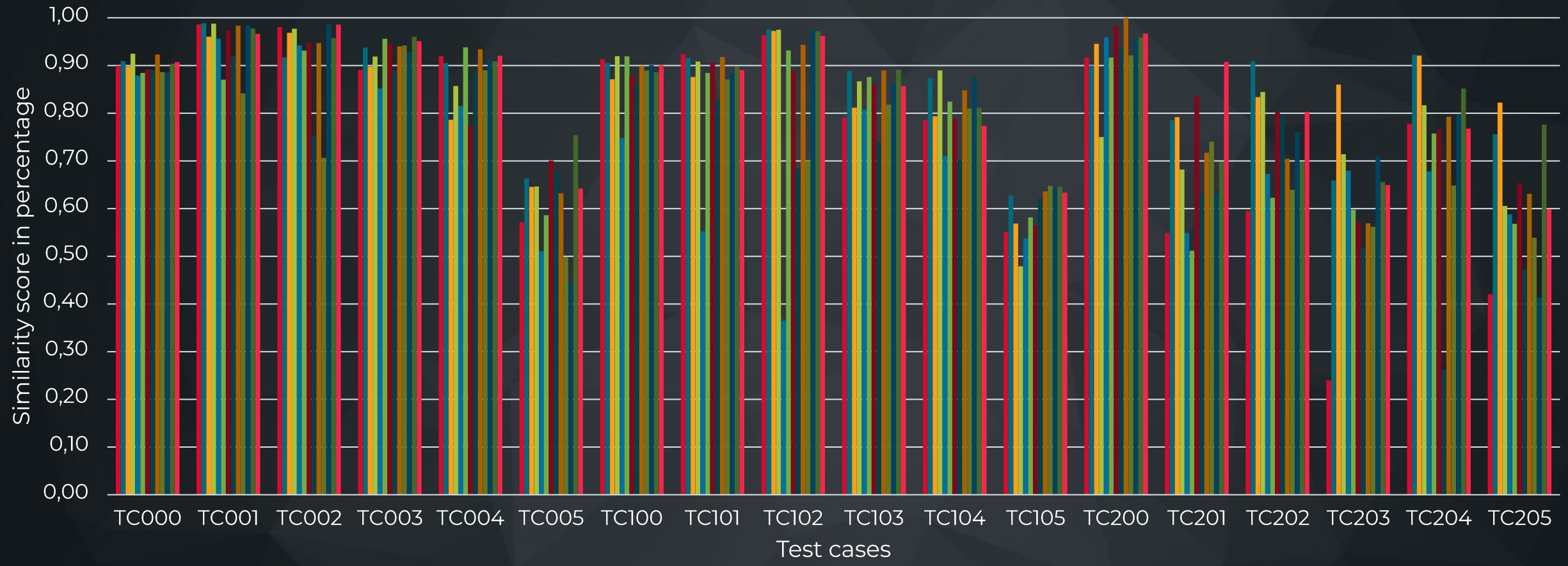
Graded Scores

6

Aggregate Results

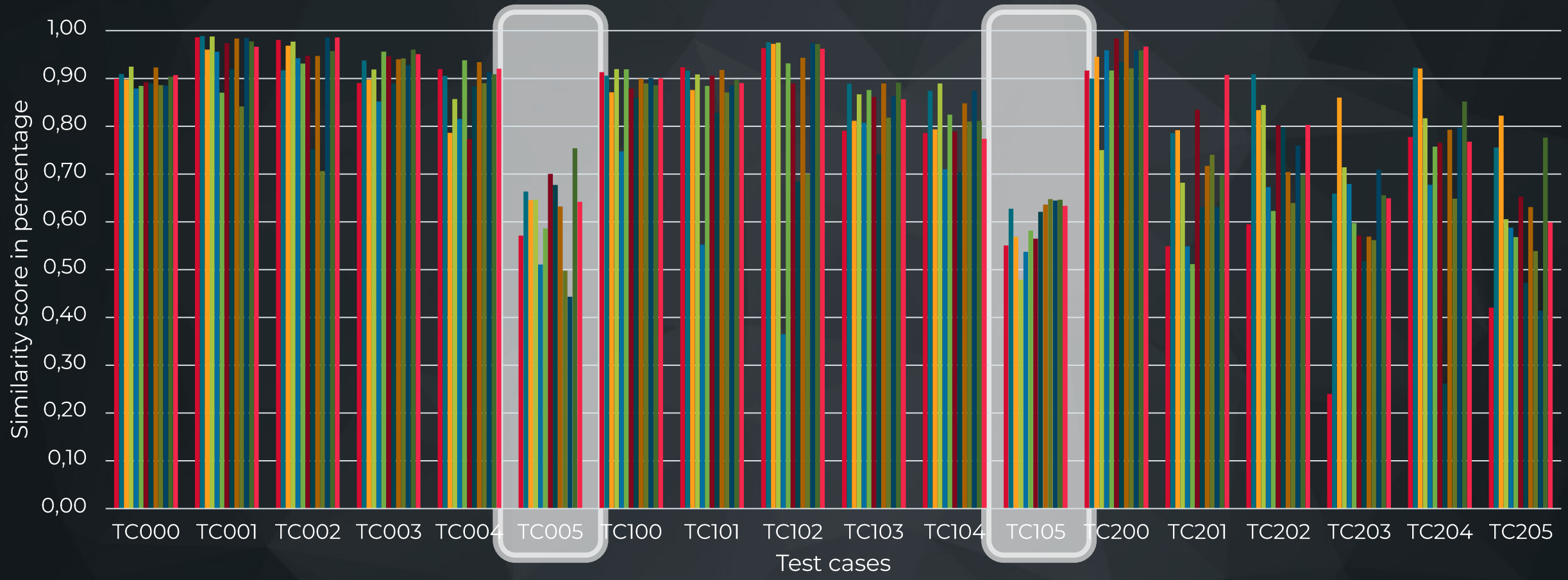
RM Bench

Our Findings



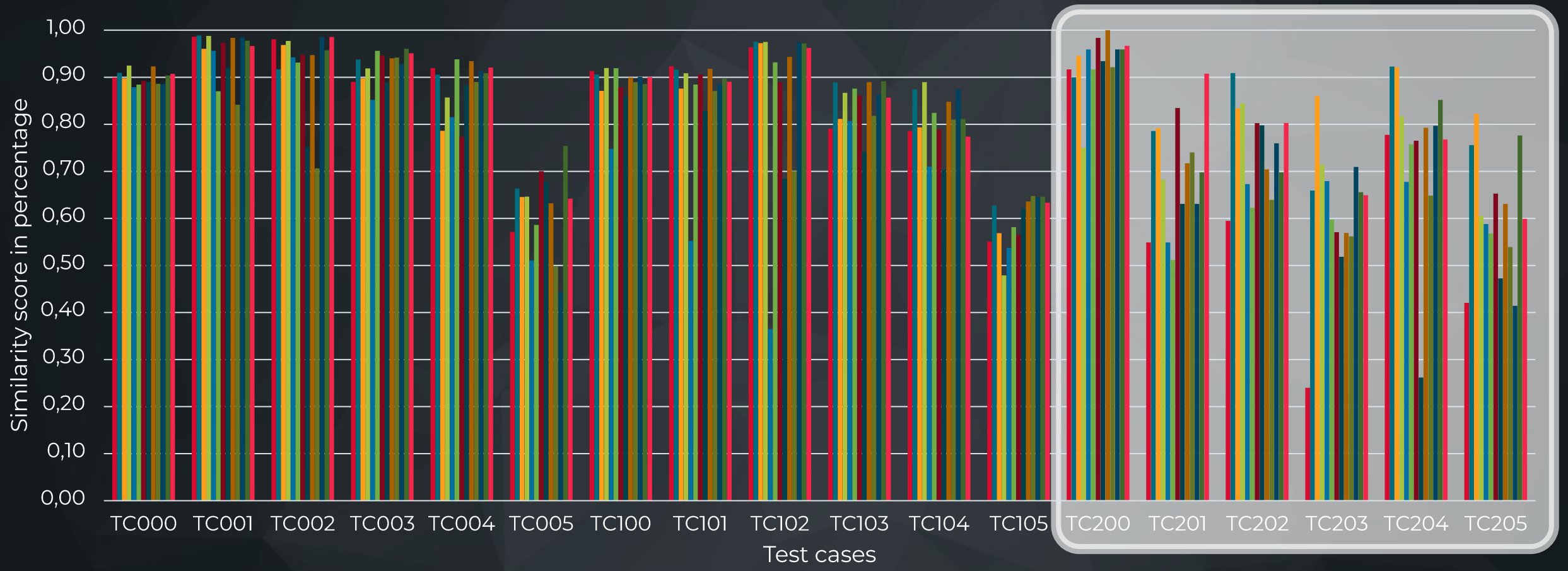
RM Bench : Our Findings

Marketplace



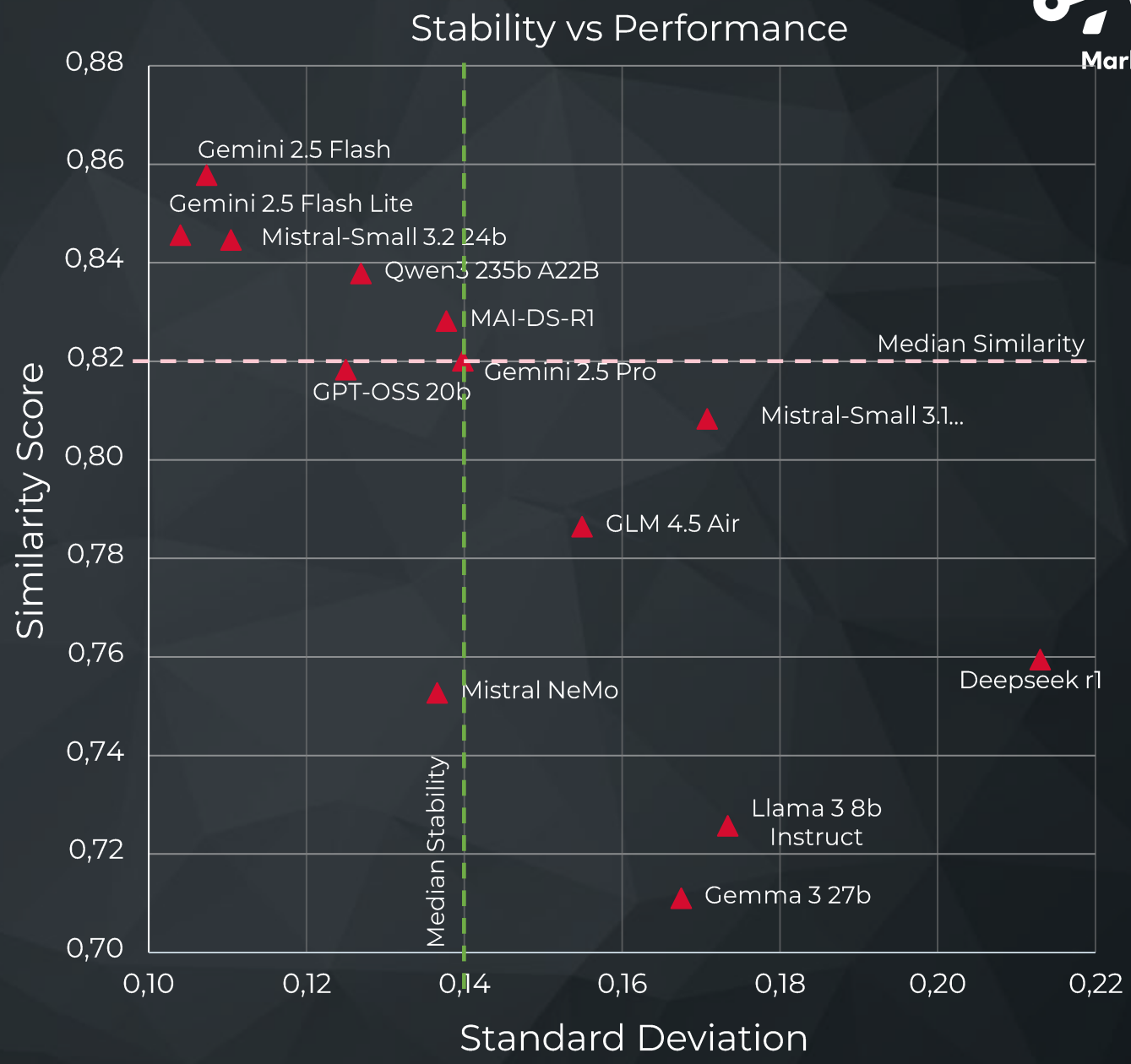
Motivation

RM Bench : Our Findings



Our Findings

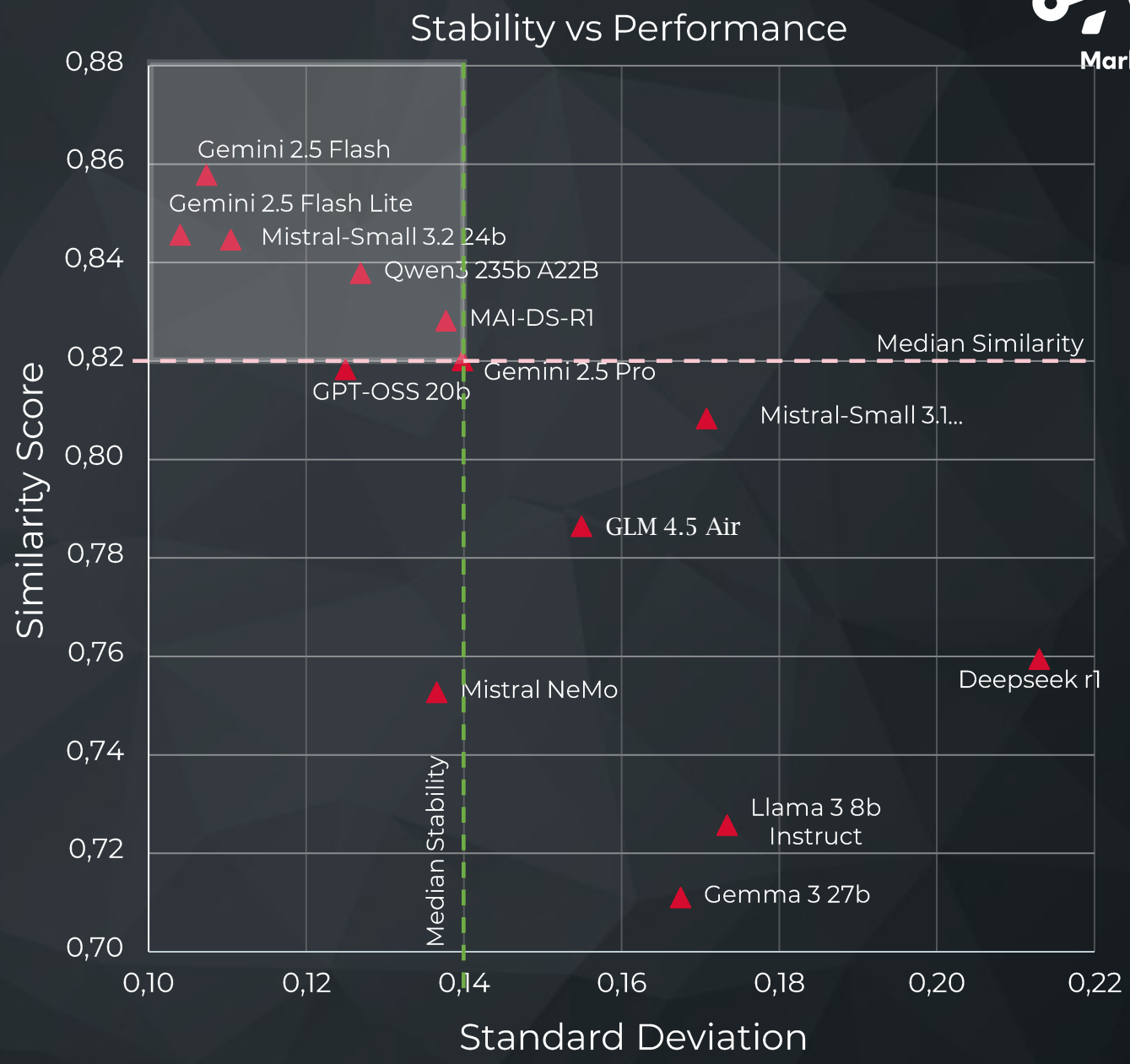
Overall Model Accuracy & Stability



Our Findings

Overall Model Accuracy & Stability

Top-performing models (e.g., Gemini 2.5 Flash, $\bar{S}$≈0.86) show high accuracy and low standard deviation (high stability).

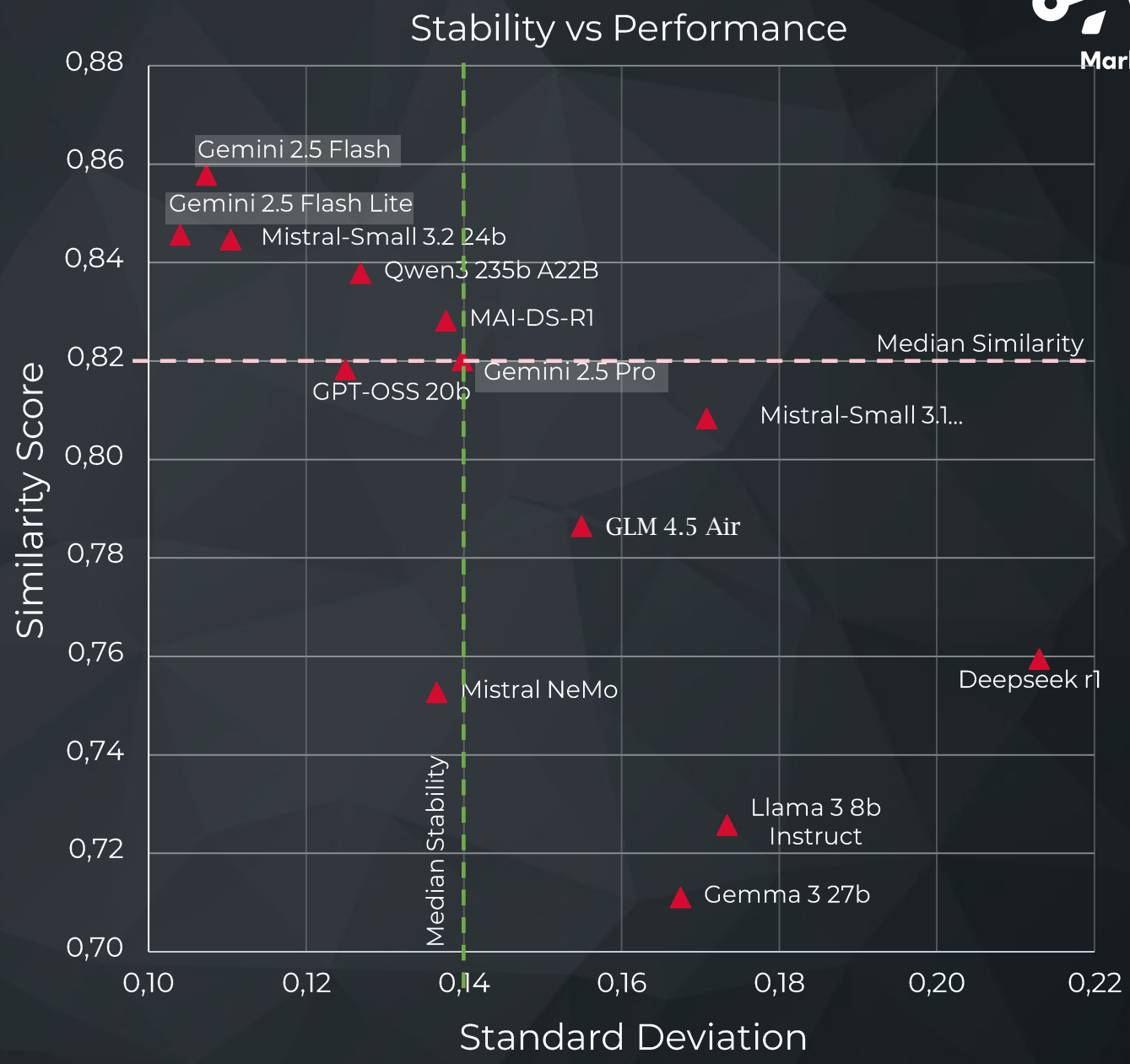


Our Findings

Overall Model Accuracy & Stability

Top-performing models (e.g., Gemini 2.5 Flash, $\bar{S}$≈0.86) show high accuracy and low standard deviation (high stability).

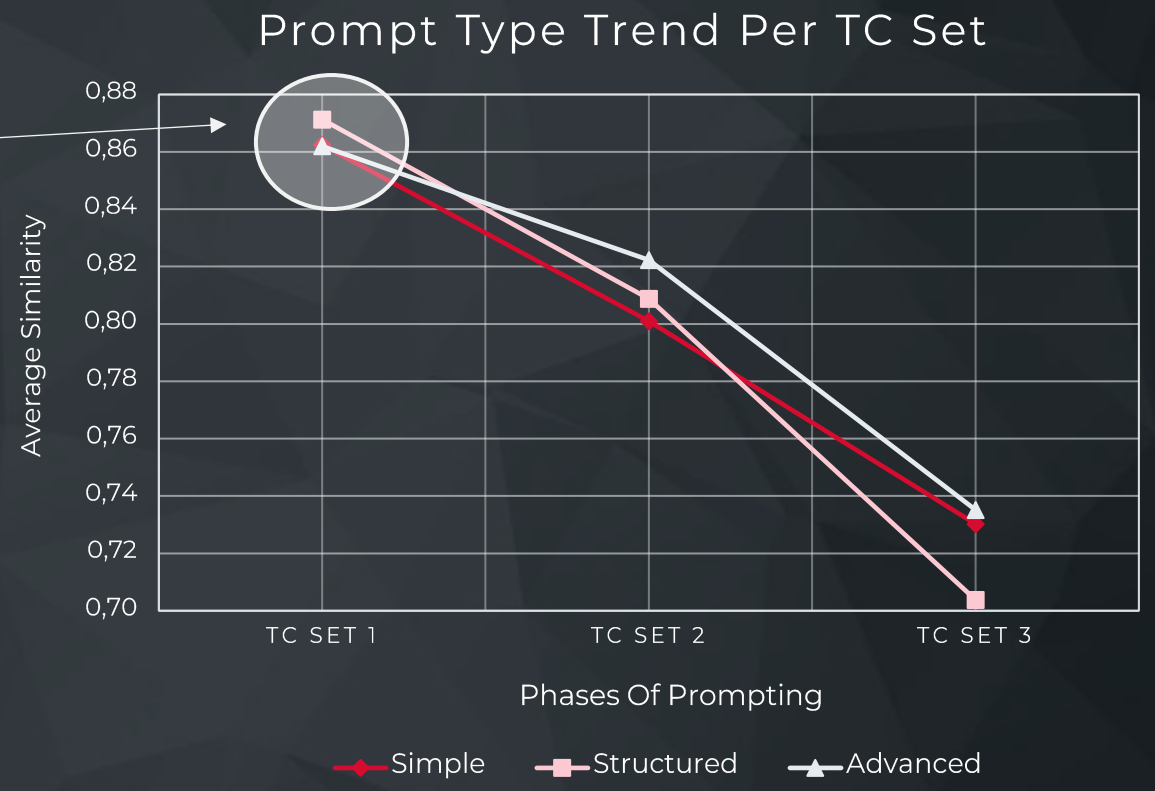
Gemini 2.5 Pro even being a better model than Gemini 2.5 Flash & Flash-Lite might be due to over-thinking.



Our Findings

Performance Across Test Case Sets

TC Set 1 (Structured Tasks) achieved highest performance ($\bar{S} \approx 0.87$).

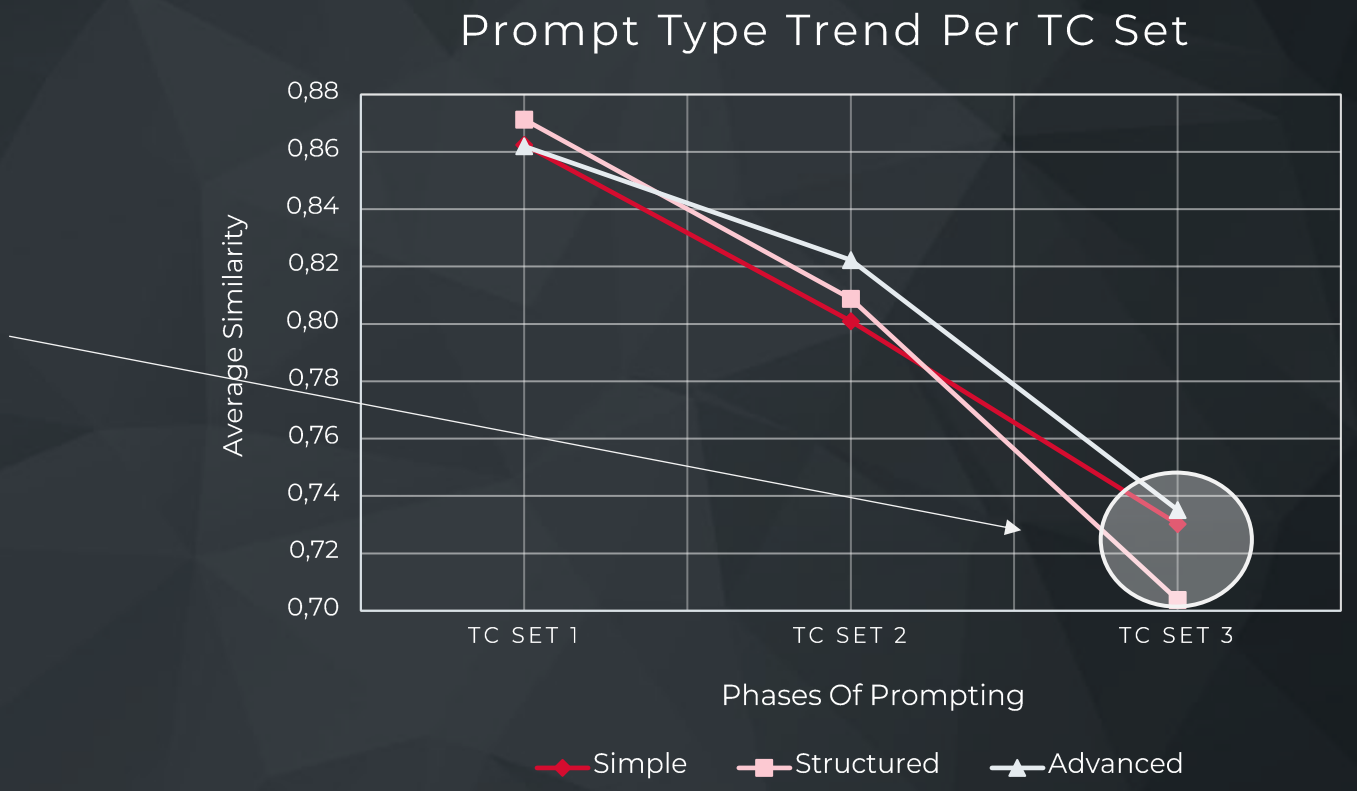


Our Findings

Performance Across Test Case Sets

TC Set 1 (Structured Tasks) achieved highest performance ($\bar{S} \approx 0.87$).

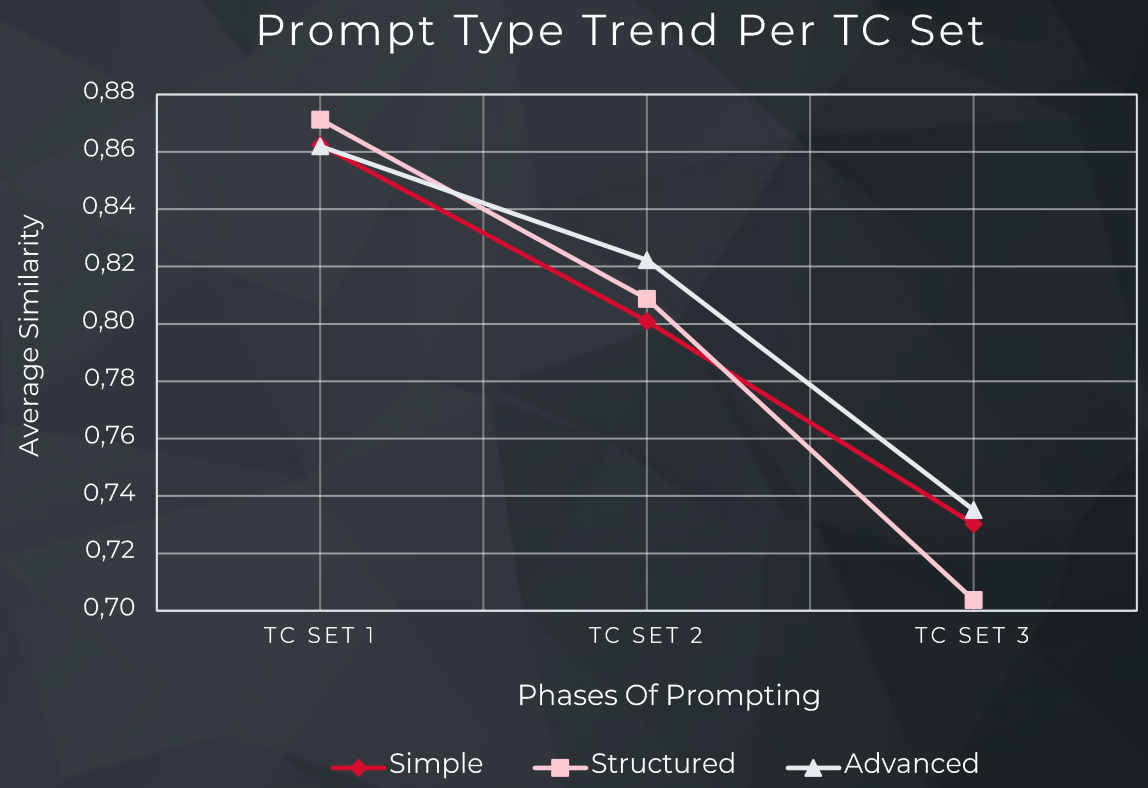
Performance dropped significantly in TC Set 3 (Semantic/Compliance Tasks) ($\bar{S} \approx 0.72$), confirming models struggle with **implicit regulatory semantics** and **contextual interpretation**.



Our Findings

Performance Across Test Case Sets

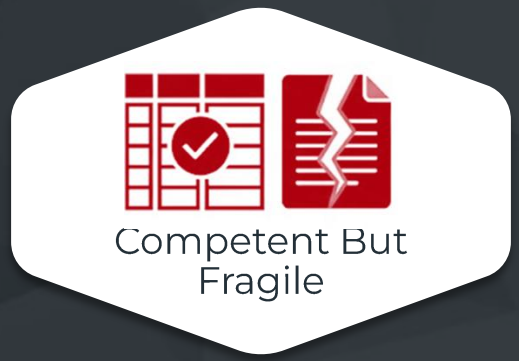
- TC Set 1 (Structured Tasks) achieved highest performance ($\bar{S} \approx 0.87$).
- Performance dropped significantly in TC Set 3 (Semantic/Compliance Tasks) ($\bar{S} \approx 0.72$), confirming models struggle with **implicit regulatory semantics** and **contextual interpretation**.
- Increasing complexity of tasks perform better with complex prompts.



RM Bench

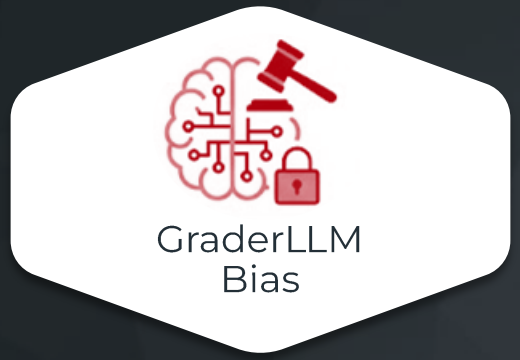
Conclusion

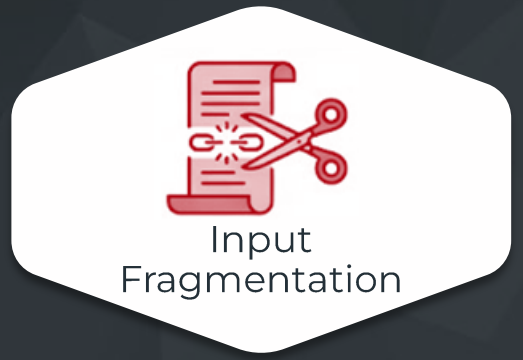
Marketplace



RM Bench

Limitations

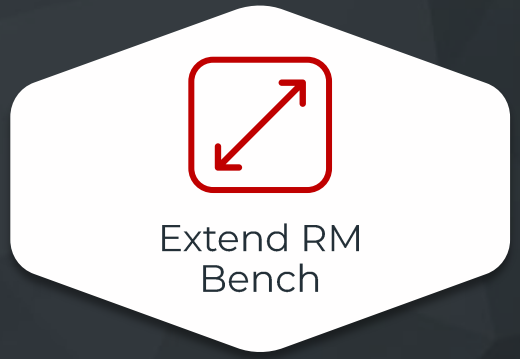

GraderLLM Bias


Input Fragmentation

Limited Scope

Future Work


Fine-tune RM Based LLMs


Extend RM Bench

Marketplace

Kontakt

nikhil.jha@ki-marktplatz.com
ruslan.bernijazov@ki-marktplatz.com