MCP Dev Summit Mumbai

Why Most MCP Tools Fail Silently And How to Measure It

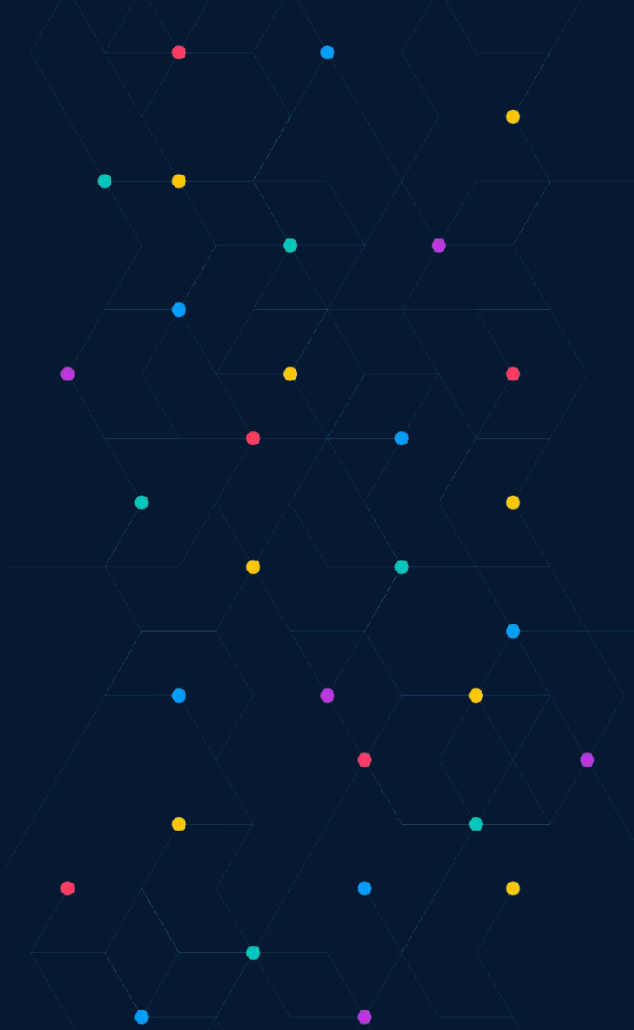
Om Shree | Shreesozo

What if

We could actually measure how good these descriptions are?

T : Tool
D : Definition
Q : Quality
S : Score

WHY SCORE TOOL DEFINITIONS?



"MCP Tool Descriptions Are Smelly!"

arXiv:2602.14878

856 tools · 103 MCP servers

- 97% have at least one quality defect
- 56% don't clearly state what the tool does
- 89% never say when to use or avoid the tool

Source: Hasan, Li, Rajbahadur, Adams, Hassan · arXiv:2602.14878

"From Docs to Descriptions"

arXiv:2602.18914

10,831 MCP servers

- **260%** more selections for well-written descriptions
- +6 % improvement for task success from rewriting descriptions alone

Source: Wang, Li, Sun, Liu, Liu, Tian · arXiv:2602.18914

For the framework to work at registry scale

It needed to be three quality signal, that is:

1. Explainable

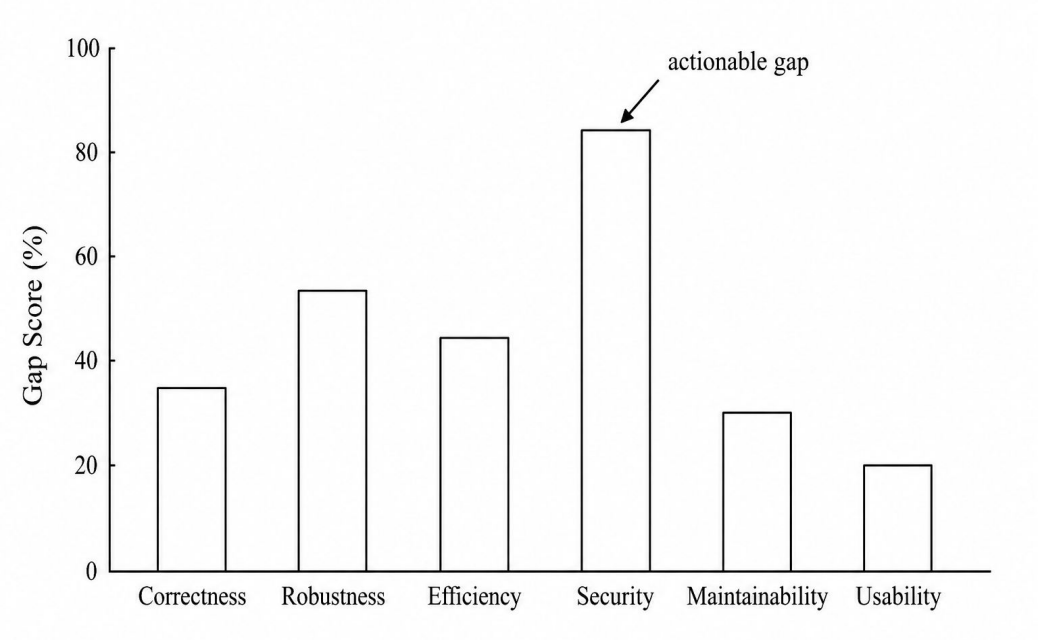


Fig 1 : Per-dimension breakdown maintainer-actionable.

For the framework to work at registry scale

It needed to be three quality signal, that is:

1. Explainable
2. Reproducible

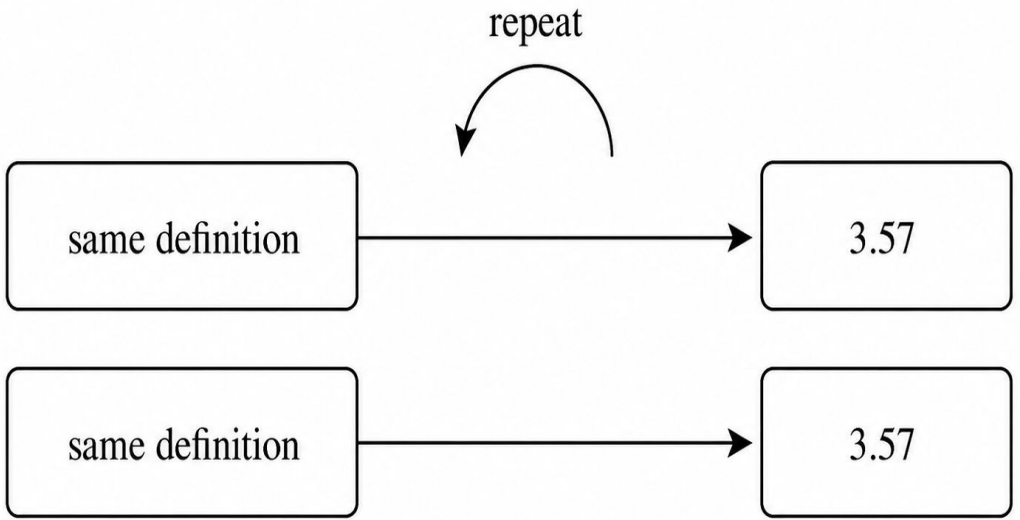


Fig 2 : Same definition always produces the same score. .

For the framework to work at registry scale

It needed to be three quality signal, that is:

1. Explainable
2. Reproducible
3. Cheap

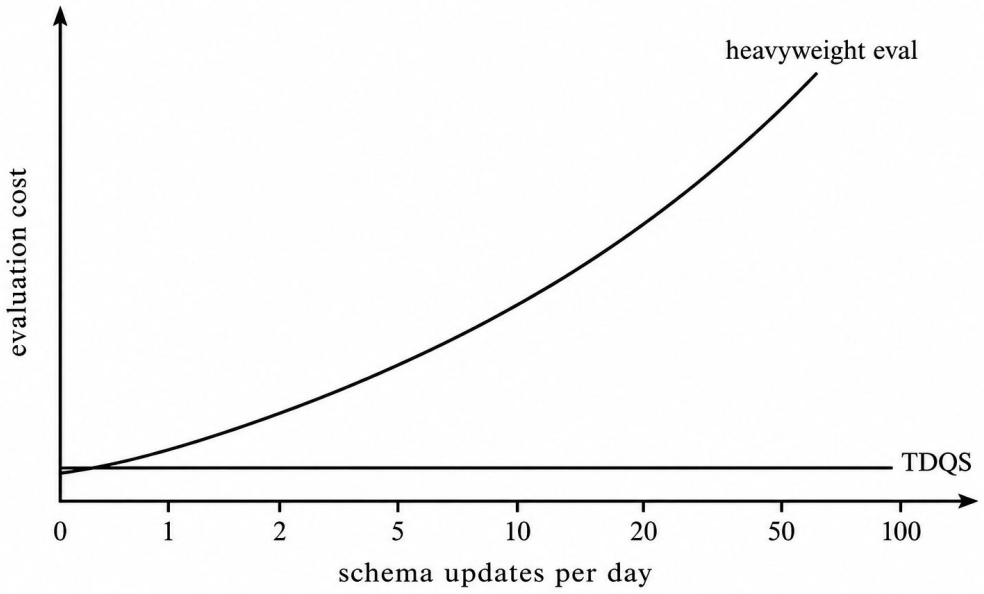
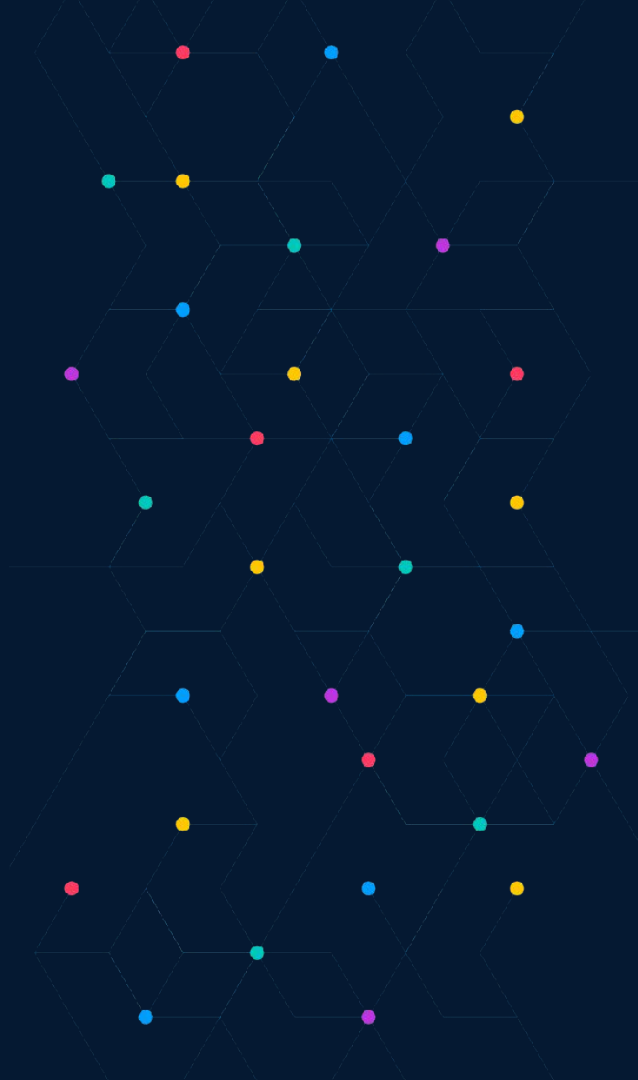


Fig 3: Thousands of updates per day - cost must stay flat.

WHAT EXACTLY GETS SCORED ???



Tool definition - not behavior.

MCP Dev Summit Mumbai

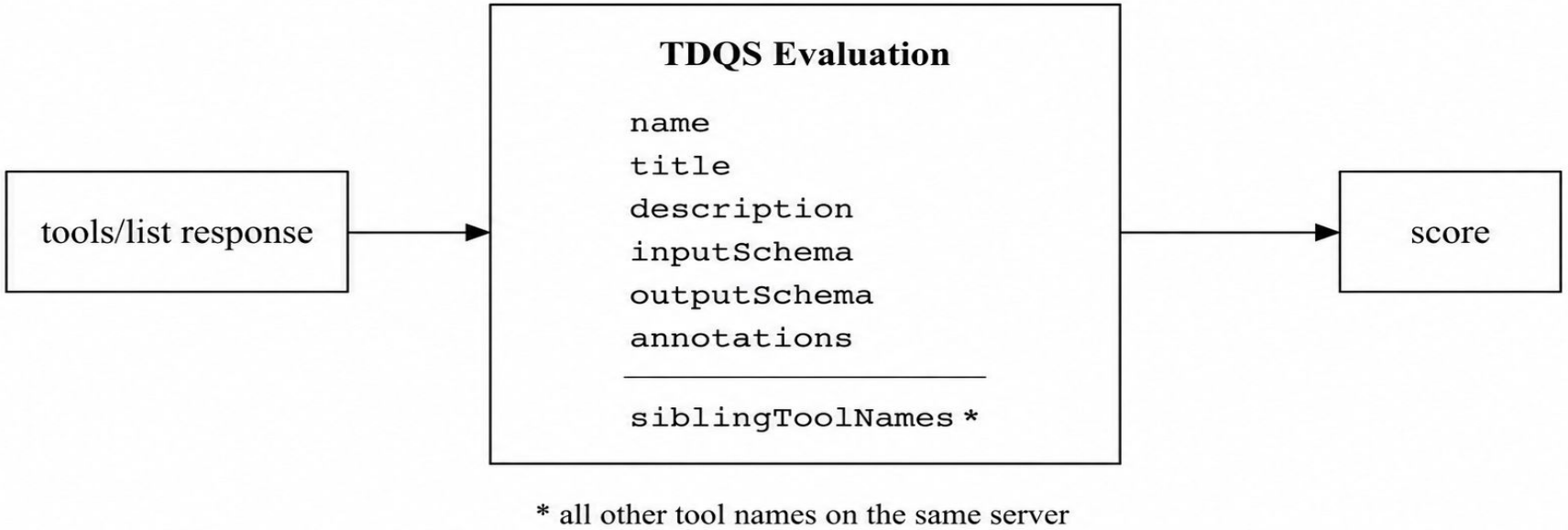
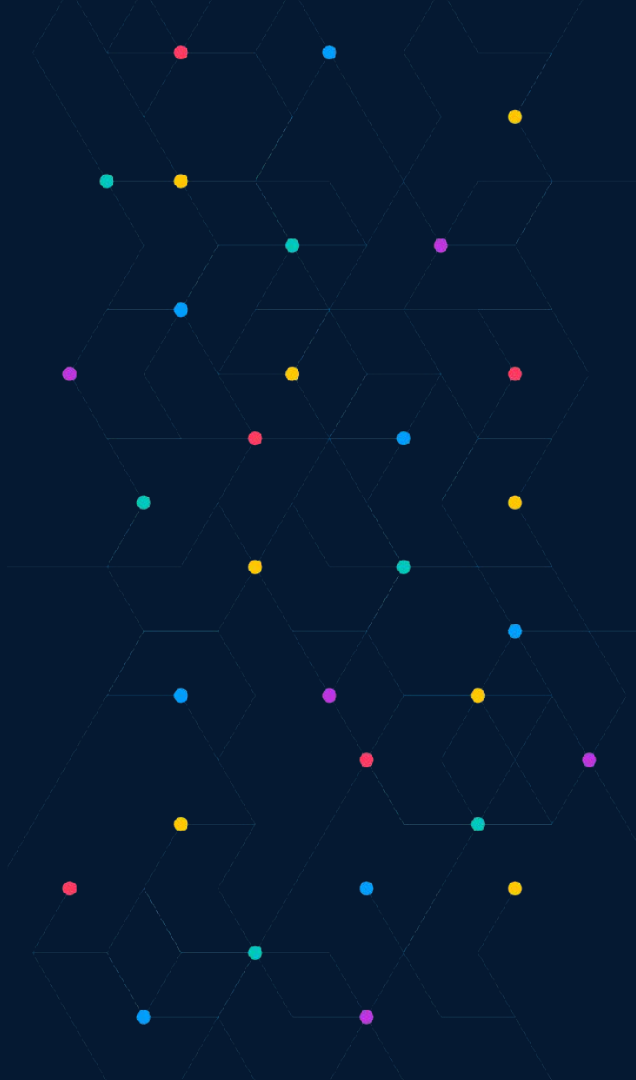
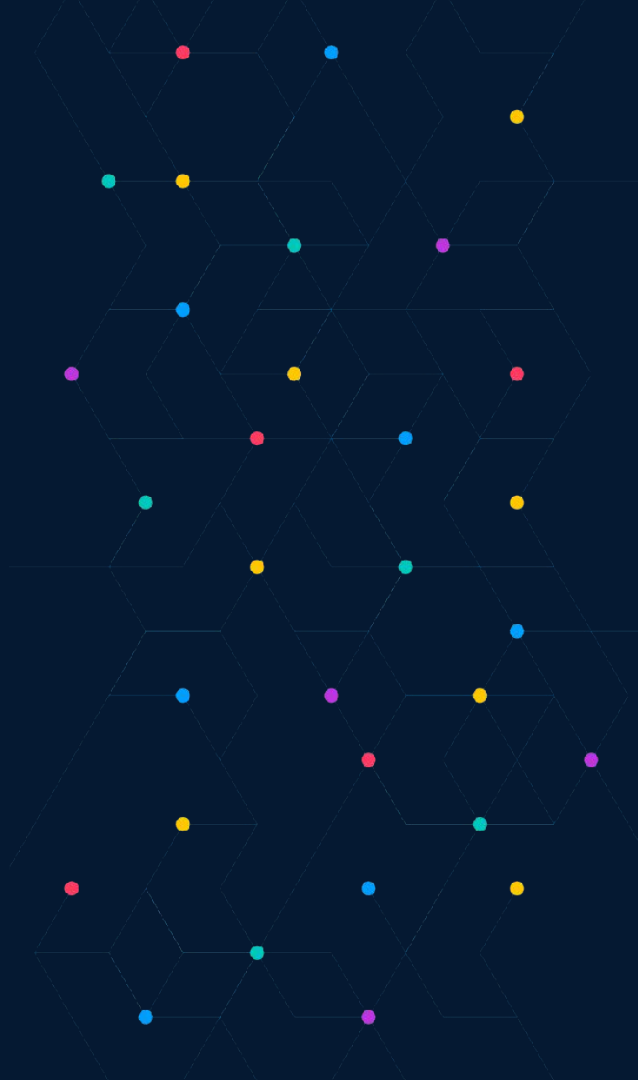


Fig 4: TDQS scores the tool definition — exactly what an MCP client receives from tools/list

Clarity is Relative
Not Absolute



THE FOUR STAGE SCORING PIPELINE



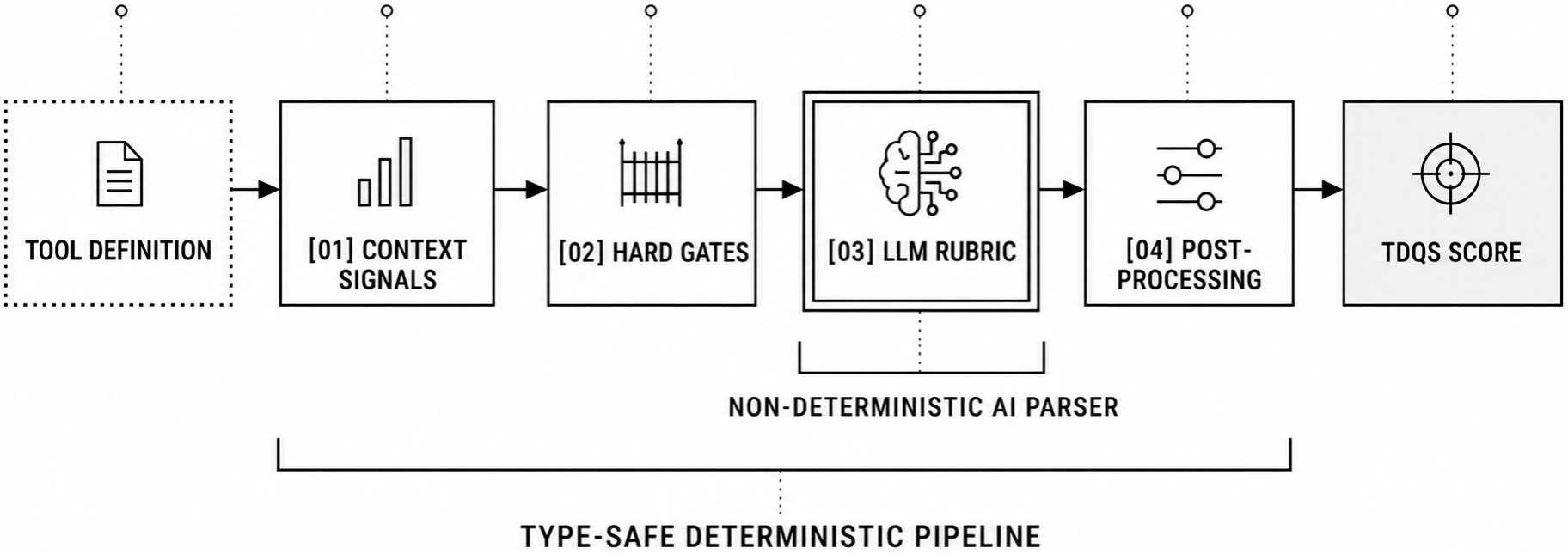
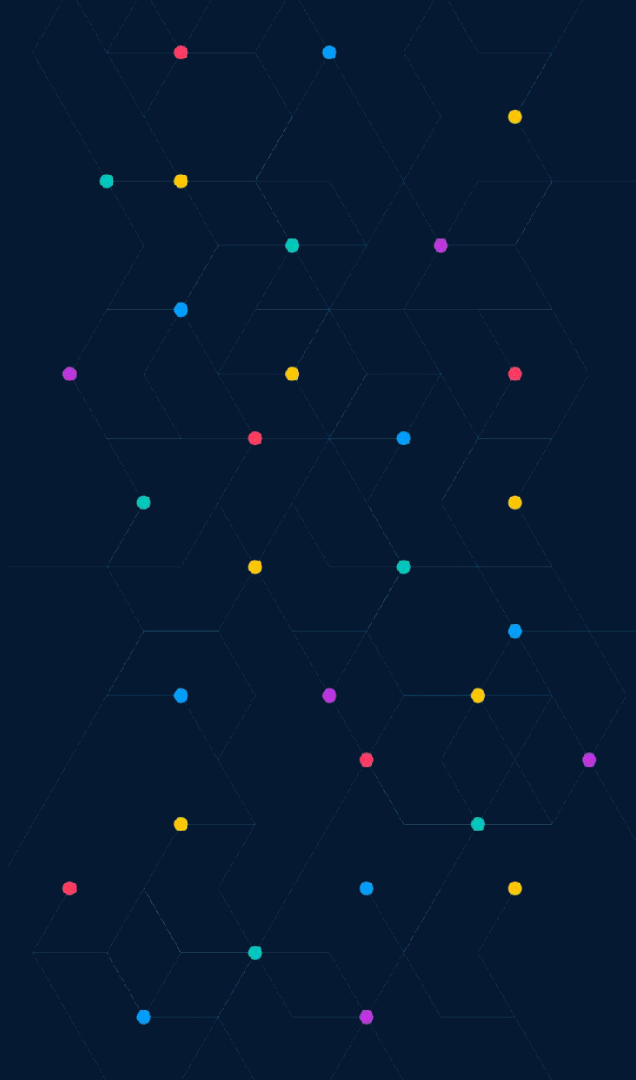


Fig 5 : The Four-Stage Assembly Line

STAGE 1: CONTEXT SIGNALS



Deterministic Code Extraction

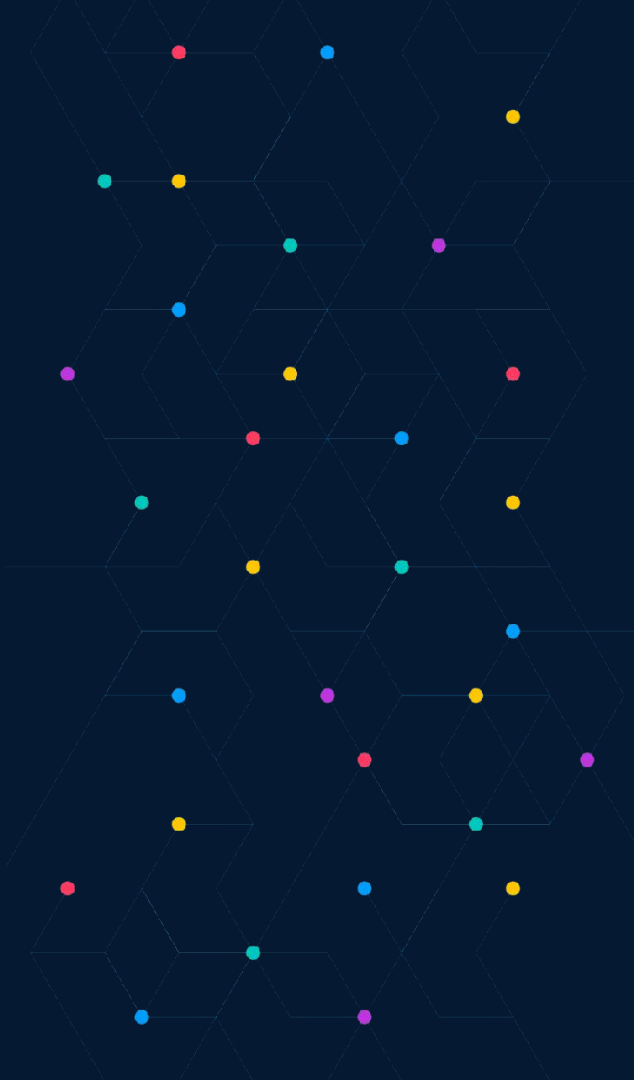
Before scoring anything, we extract the raw facts — param counts, description coverage, a SHA-256 fingerprint — straight from the tool definition.

"We read the tool definition and extract cold hard facts before any judgment is made."

Analogy

Like a doctor checking your height, weight, and blood pressure before the actual diagnosis begins.

STAGE 2: HARD GATES



Automated Short-Circuits

Two automatic checks that short-circuit the whole pipeline before the LLM even sees the tool.

"Two automatic checks that disqualify a tool before the LLM wastes a single token on it."

Analogy

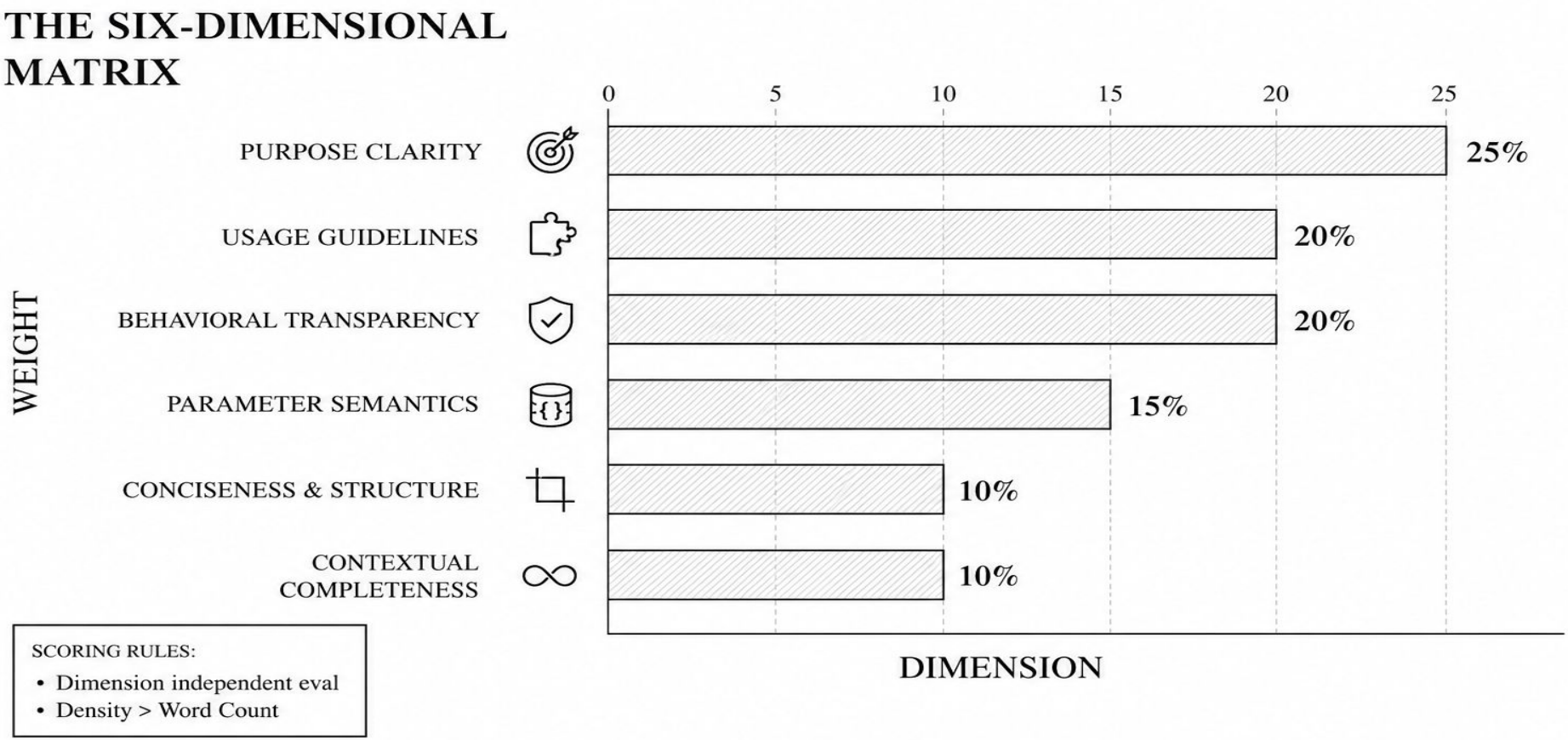
Like a bouncer at the door — no ID, no entry. No description, no score.

STAGE 3: LLM RUBRIC EVALUATION

LLM Rubric Eval

An LLM scores the tool across six dimensions, with a 2–3 sentence justification required for every score.

"An LLM reads the tool and scores it across six quality dimensions, one by one."



MCP Dev Summit Mumbai

Fig 6 : The Six-Dimensional Matrix

Analogy

Like a judge scoring a gymnastics routine, not just pass or fail, but points for every element.

STAGE 4: POST-PROCESSING

Sanitization & Feature Locking

The pipeline clamps, validates, and converts raw scores into a final TDQS number and a tier.

"An LLM reads the tool and scores it across six quality dimensions, one by one."

Analogy

Like a teacher totalling marks, applying grade boundaries, and printing your report card.

Computing the Final Score

MCP Dev Summit Mumbai

TOP MODULE

$$\text{TDQS} = (P \times 0.25) + (U \times 0.20) + (B \times 0.20) + (\text{S} \times 0.15) + (C \times 0.10) + (X \times 0.10)$$

P=Purpose, U=Usage, B=Transparency, S=Semantics, C=Conciseness, X=Completeness

MIDDLE MODULE **WORKED EXAMPLE: TIER C INTERFACE**

Input Scores: purpose=4, guidelines=2, transparency=2, params=3, conciseness=4, completeness=2

Analytical Resolution:

$$(4 \times 0.25) + (2 \times 0.20) + (2 \times 0.20) + (3 \times 0.15) + (4 \times 0.10) + (2 \times 0.10) = \mathbf{2.85}$$

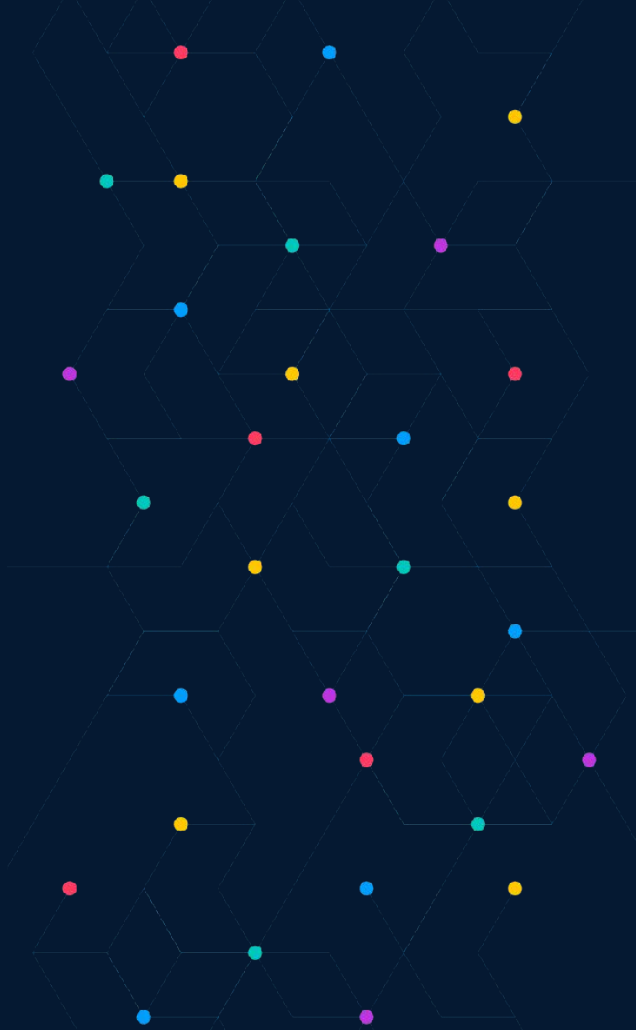
↓

➔ TDQS 2.9

BOTTOM MODULE **PERFORMANCE TIER CLASSIFICATION**

TIER A (≥ 3.5)	TIER B (≥ 3.0)	TIER C (≥ 2.0)	TIER D (≥ 1.0)
GENUINELY HELPFUL	**ADEQUATE (PASSING BAR)**	**CLEAR GAPS**	**SEVERELY DEFICIENT**

TDQS ACROSS THE REGISTRY



73.5% Clear The Passing Bar !

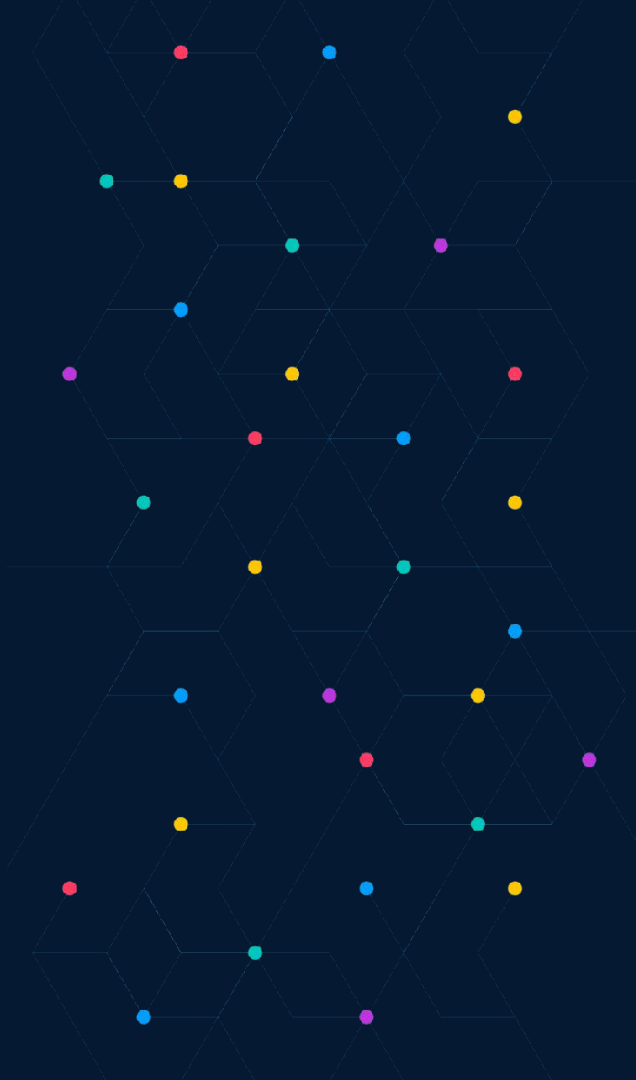
Tier	Tools	Share
A (≥ 3.5)	**195,872**	**54.1%**
B (≥ 3.0)	**70,186**	**19.4%**
C (≥ 2.0)	86,303	23.8%
D (≥ 1.0)	9,865	2.7%

Per-dimension breakdown maintainer-actionable.

The Composite Hides where things break.

Dimension	Weight	Mean	Smell rate (< 3)
Purpose Clarity	25%	4.47	4.0%
Conciseness & Structure	10%	4.47	3.8%
Contextual Completeness	10%	3.23	32.8%
Parameter Semantics	15%	3.19	14.0%
Usage Guidelines	20%	3.03	44.5%
Behavioral Transparency	20%	2.90	46.1%

IMPROVING YOUR SCORE



The TDQS Optimization Blueprint

1. One sentence. Verb + resource + scope + sibling diff

Analogy : One sentence. Verb + resource + scope

Like a good job listing: tells you the role, the team, and why it's not the same as the other opening.

The TDQS Optimization Blueprint

1. One sentence. Verb + resource + scope + sibling diff
2. Say when NOT to use it. Name the alternative.

Analogy : Say when NOT to use it. Name the alternative.

Like a medicine label that says 'not for children under 12, use X instead.' The warning is as valuable as the drug.

The TDQS Optimization Blueprint

1. One sentence. Verb + resource + scope + sibling diff
2. Say when NOT to use it. Name the alternative.
3. Declare annotations. readOnlyHint, destructiveHint.

Analogy : Declare annotations. readOnlyHint, destructiveHint.

Like a wet floor sign in a corridor: it doesn't stop you from walking, but it tells you exactly what you're stepping into.

The TDQS Optimization Blueprint

1. One sentence. Verb + resource + scope + sibling diff
2. Say when NOT to use it. Name the alternative.
3. Declare annotations. readOnlyHint, destructiveHint.
4. Never contradict your annotations.

Analogy : Never contradict your annotations.

Like a fire exit sign pointing left, but the door is on the right. The sign exists, but it kills you.

Final Goal


Tool Definition Quality
A
5/5.0
Behavior
5/5
Conciseness
5/5
Completeness
5/5
Parameters
5/5
Purpose
5/5
Usage Guidelines
5/5

How Glama runs this on 228,000 tools.

1. Sweep every 5 min · batches of 100
2. inputHash match = skip · no LLM call
3. One changed tool = one re-score
4. Every LLM response schema-validated · retried on mismatch

Source :

https://glama.ai/blog/2026-04-03-tool-definition-quality-score-tdqs

glama-ai/**tool-definition-quality-score**

An open framework for scoring how well an MCP tool definition communicates to an AI agent – the Tool Definition...

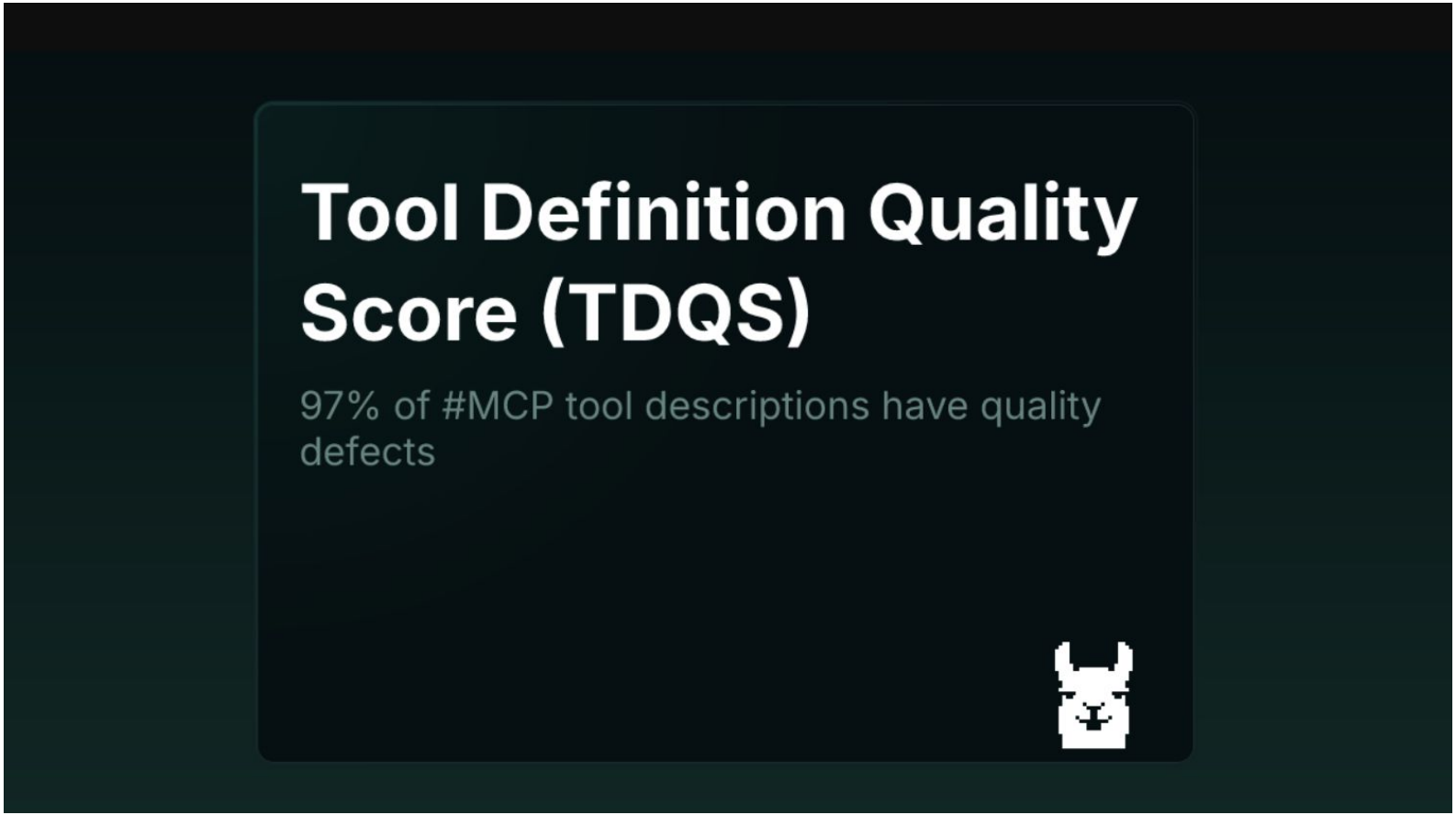
1 Contributor

0 Issues

3 Stars

1 Fork

LINK: https://github.com/glama-ai/tool-definition-quality-score



LINK: https://glama.ai/blog/2026-04-03-tool-definition-quality-score-tdqs

Why Most MCP Tools Fail Silently, And How to Measure It

MCP Dev Summit Mumbai

Om Shree

Technical Evangelist | Simplifying Complex AI & Agent Workflows for Devs | Building Shreesozo

MCP Dev Summit Mumbai