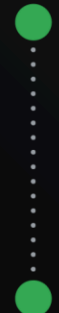


Language Models Like Structure Too



Rory Pilgrim
Product Manager, Google Research



Central Claim

Adopting a standardized, structured format for industrial data will accelerate the speed and impact AI can have to improve your workflows

Two main points

1. AI works better with structured data

Agentic workflows enable AI to take advantage of structured data

This leads to increased reliability, reduced cost and reduced latency

2. Standardizing that data focuses the community

Creating a standard for your domain with challenging evals incentivizes pre-training and hill climbing by AI labs

This improves out-of-the-box performance for models to both utilize the structured data and also write to it

Two main points

1. AI works better with structured data

Agentic workflows enable AI to take advantage of structured data

This leads to increased reliability, reduced cost and reduced latency

2. Standardizing that data focuses the community

Creating a standard for your domain with challenging evals incentivizes pre-training and hill climbing by AI labs

This improves out-of-the-box performance for models to both utilize the structured data and also write to it

Larger context degrades AI performance

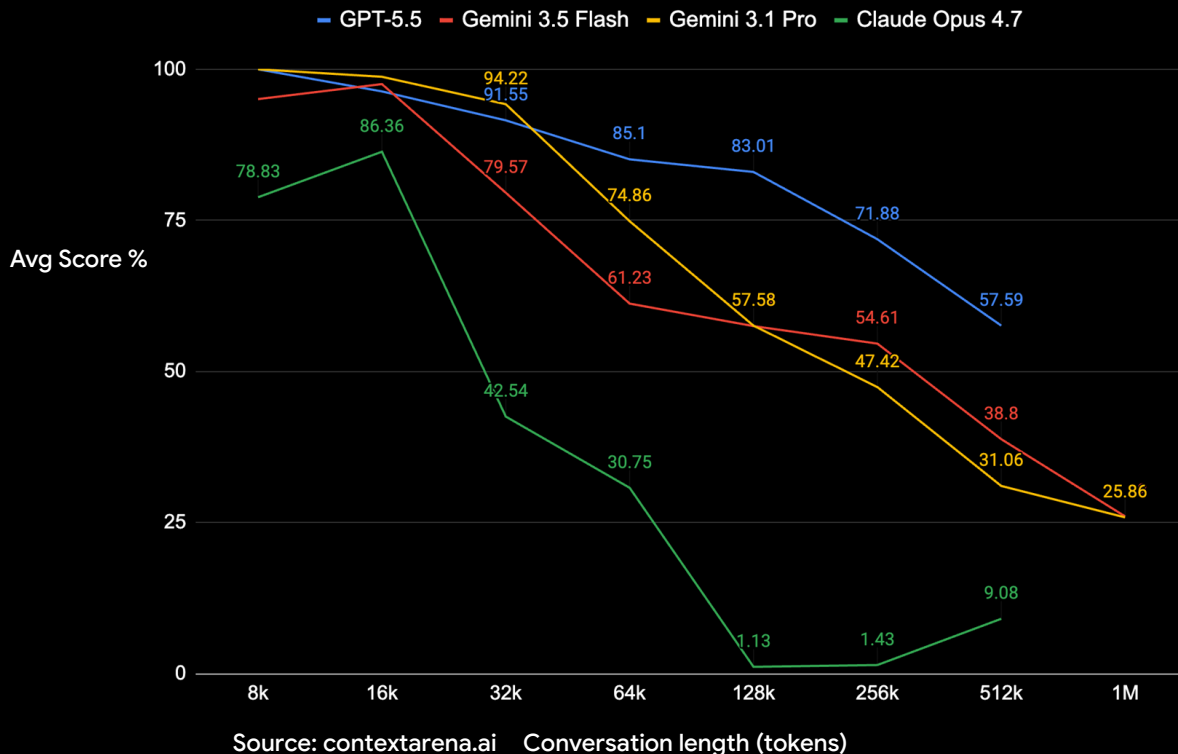
Multi-Round Coreference Resolution (MRCR) is an AI eval measures how LLMs retrieve information across long conversations.

Includes distractor information (“needles”) that the model must ignore

Example: “Add the string “AKJSs89sal” to the 2nd poem about penguins.” Where the whole massive conversation involves 8 interspersed conversation for need=8

[MRCR Paper](#)

MRCR V2 Performance at difference context bins (8 Needle)



Agentic Advent: how our use of LLMs has changed in the last 18 months

2022-2024: Wave 1 LLMs

Conversation progresses 1 turn at a time.

Prepend previous turns as context each time.

For users= Lots of copying and pasting

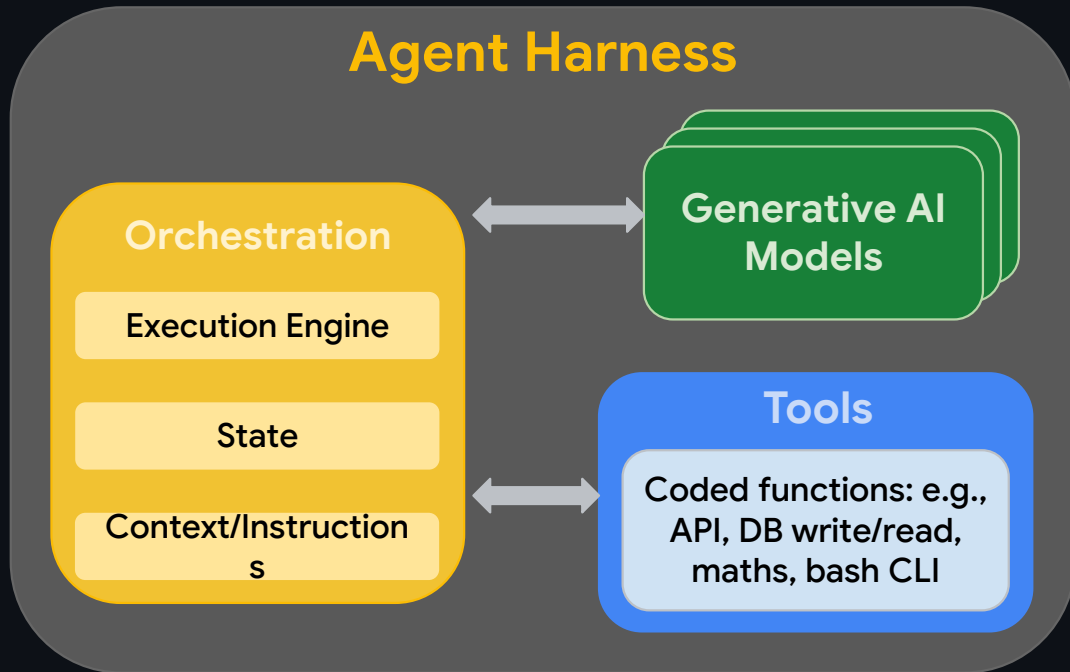


2025-now: Agentic AI

LLM in a Loop with the ability to execute code i.e., “tools”

What is an agent?

“Autonomous system that combines GenAI outputs with pre-defined coded functions (“tools”), to achieve some purpose”



Rerunning MRCR Agentically

We can make MRCR into an agentic eval. Instead of feeding the conversation into context window:

1. Store the conversation in a database
2. Give the agent a SQL tool it can use to query the conversation

Conversation Database		
turn	speaker	text
1	user	"Please...
2	assistant	"Sure, happy to

Result: Gemini 3.5 Flash scores **100%** performance at 1M token bin vs **26%** in long context

User Query:

Prepend ZnDFz0Qxoz to the 5th (1-indexed) poem about presidents. Do not include any other text.

Tool Call:

Request:

```
SELECT turn, text FROM conversation_turns
WHERE speaker = 'user'
AND text LIKE '%poem%'
AND text LIKE '%president%'
ORDER BY turn
```

Result: Turn #s 6, 10, 12, 16, 30, 38, 48, 80

Tool Call:

Request:

```
SELECT text FROM conversation_turns
WHERE turn = 31
```

Result: "Presidents past, in halls of time..."

Final Output:

ZnDFz0QxozPresidents past, in halls of time, Stewards of peace...

A more realistic example: Health Records

Fast Healthcare Interoperability Resources (FHIR)

A standardized, JSON schema for healthcare data used for interoperability of electronic health records across hospitals.

Unexpected bonus: Enables LLMs to natively understand and efficiently navigate complex and large clinical records without context bloat.

- `patient.fhir.json`

```
{
  "resourceType": "Patient",
  "id": "example-001",
  "name": [{
    "family": "Smith",
    "given": ["John"]
  }],
  "birthDate": "1980-04-15",
  "gender": "male",
  "address": [{
    "city": "London",
    "country": "UK"
  }]
}
```

FHIR Experiment: Structured Agent vs Long Context

Based on 1.4M-token FHIR patient: "What is the maximum lifetime heart rate for this patient?" record

Approach	Agent	Long Context
Method	Agent calls searchResources for Observations, receives a clean array of 311 filtered JSON objects	Splits 1.4M token patient record into chunks, asks model to scan and merge
Result	✅ Correctly identifies the global maximum [91]	❌ "Lost in the Middle" — chunks hallucinate local maximums [88, 89] and fail to merge
Latency	22s	57s
Cost	\$0.20 (101K tokens)	\$5.51 (1.38M tokens)

Structure is not just cheaper and faster — it is required for global reasoning over massive histories.

How this relates to industrial applications

For AI Agents to be maximally effective they need a unified data structure to query, and retrieve information across your entire plant to complete tasks across engineering lifecycle:



The diagram consists of three dark blue rounded rectangular boxes arranged horizontally. Each box is topped with a thin horizontal line of a different color: blue for Design, orange for Operations, and green for Maintenance. The text inside each box is white and bold.

Design

Operations

Maintenance

Recap: why structure matters for AI

Context Limits: LLMs have a limited context window and perform worse as you fill that context window.

Agent Solution: Agents manage this, they allow LLMs to retrieve the right data for the task at hand.

Core Requirement:

For agents to be effective the data they operate on must be structured.

Two main points

1. AI works better with structured data

Agentic workflows enable AI to take advantage of structured data

This leads to increased reliability, reduced cost and reduced latency

2. Standardizing that data focuses the community

Creating a standard for your domain with challenging evals incentivizes pre-training and hill climbing by AI labs

This improves out-of-the-box performance for models to both utilize the structured data and also write to it

Training an LLM for your use cases is expensive, create a benchmark to persuade someone else to do it for you

Capture Attention

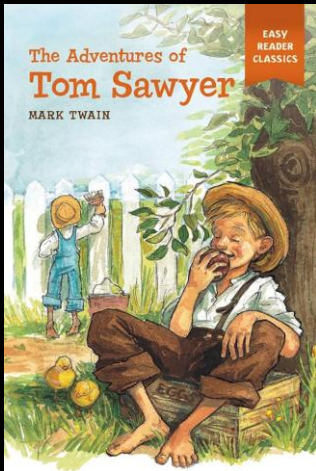
Launch an open standard paired with a diverse, unsaturated benchmark spanning real-world use cases.

Force Canonical Adoption

Become the definitive industry benchmark. This obliges frontier labs to invest \$XXM dollars targeting these specific tasks during training.

Reap the Dividends

Models arrive highly optimized out-of-the-box. Your specific applications become vastly more performative and significantly cheaper to run.



Sample of leading benchmarks

Eval	Developed by	Description
Vals Finance Agent	Vals, private AI eval company	Has become the standard benchmark reported by companies for showcasing financial performance
OSworld	Academic + Sales Force	Leading evaluation for interactive Computer environment for
Terminal Bench	Academic + Industry	Leading benchmark for terminal usage

Sharing knowledge, swapping skills

What are agent skills?

Skills are reusable, domain-specific instructions that teach agents workflows without bloating their context.

They include detailed instructions, links and scripts.

Context efficient: They are only pulled into context as the agent requires them.

A standard allows joint development of skills

A standardized data structure enables users to share skills across industries and companies and to collaborate together to improve performance.

“Many hands make light work”

Skills could address common use cases, steer agents on blind spots or provide safety scripts for work integrity.

Central Claim

Adopting a standardized, structured format for industrial data will accelerate the speed and impact AI can have to improve your workflows