sprinklr

28TH ICPR 2026 LYON, FRANCE ORAL PRESENTATION

CXMArena

A Unified Dataset to Benchmark Performance in Realistic CXM Scenarios

Raghav Garg · **Kapil Sharma** · **Karan Gupta**

Sprinklr

Benchmarking LLMs on the operational tasks that run a contact center

THE PROBLEM

LLMs could transform the contact center — but can we trust them?

01 **Huge real-world stake.** Customer Experience Management (CXM) — running contact centers — is among the biggest enterprise uses of LLMs, over chat, voice, and email.

02 **But we can't evaluate them well.** Real interaction data is locked behind privacy walls — genuine transcripts are scarce.

03 **And existing benchmarks aren't realistic.** They test conversational fluency, not whether the model can actually do the job.

WHAT REAL BENCHMARKS MISS

- ✕ Deep knowledge-base integration
- ✕ Real-world noise (ASR, IVR, typos)
- ✕ Operational tasks beyond dialogue
- ✕ Closed-domain answer provenance

THE GAP

Prior benchmarks test conversation — not operations

EXISTING DATASETS

- **SQuAD / HotpotQA / NQ**
  open-domain Wikipedia, no closed-KB provenance
- **Bitext / MASSIVE**
  message-level intents, only ~27 classes
- **QAConv / CoQA**
  dialogue understanding, not quality compliance
- **MTRAG / BFCL**
  no unified KB-grounded multi-turn + tool use

vs

WHAT CXM OPERATIONS ACTUALLY NEED

- ✓ Retrieval inside a closed, domain-specific KB
- ✓ Conversation-level intents over 100+ nuanced classes
- ✓ Adherence to explicit quality & compliance standards
- ✓ Multi-turn RAG with grounded passages AND tool calls

Each of those four needs becomes a task in CXMArena.

OUR CONTRIBUTION

CXMArena: a realistic, unified CXM benchmark

A large-scale **synthetic** benchmark that sidesteps privacy, injects **real-world noise**, and **co-generates** KBs with the conversations grounded in them.

01

The Dataset

Five validated benchmarks across the CXM tasks prior work neglects — plus the full underlying conversations & KB with metadata for downstream research.

02

The Pipeline

A scalable, LLM-powered generator: persona-driven KBs & dialogues, controlled noise informed by domain experts, and automated validation — language- and domain-agnostic.

THE BENCHMARK

Five operational tasks that keep a contact center running

01

KB Refinement

Flag redundant or contradictory KB article pairs

keep knowledge correct

02

Intent Prediction

Classify the whole conversation over 100+ intents

know why they called

03

Agent Quality Adherence

Audit a transcript against quality & compliance rules

assure service quality

04

Article Search

Retrieve the right KB article for a query

find the answer

05

Multi-turn RAG + Tools

Predict the next KB passage or tool call

assist in real time

Two families of work: knowledge & retrieval, and conversation analytics.

HOW WE BUILD IT

One pipeline generates it all — KBs, conversations, and labels

1

Brand & KB Generation

Information + Issue KBs from a topic tree

2

Conversation Simulation

Personas, KB-grounded turns, tool calls

3

Controlled Noise + Validation

ASR, IVR, formatting noise + auto-validation

4

5 Benchmark Datasets

Labels from traceable ground-truth links

Co-generated by design. KBs and conversations are created together, so every agent turn traces back to a specific KB passage or tool call — verifiable ground truth for retrieval, RAG, and quality tasks.

WHY IT'S REALISTIC

Realism comes from controlled, expert-informed noise

Clean generated data isn't realistic. We perturb it the way real contact-center data is messy:

ASR errors

phonetic errors on voice channels

IVR entry

menu navigation before the agent

Fragmentation

split, hesitant, interrupted speech

Formatting

HTML, Markdown, tables, PDFs

Acronyms

terms swapped for abbreviations

✓ Plus automated validation on every task — so the dataset stays realistic and trustworthy.

SCALE

The dataset at a glance

Conversations

1,994

simulated agent–customer dialogues

KB articles

3,168

1,743 informational · 1,425 issue-resolution

Callable tools

150

APIs an agent can invoke mid-conversation

Intent classes

208

in the largest of three taxonomies

Labeled article pairs

811

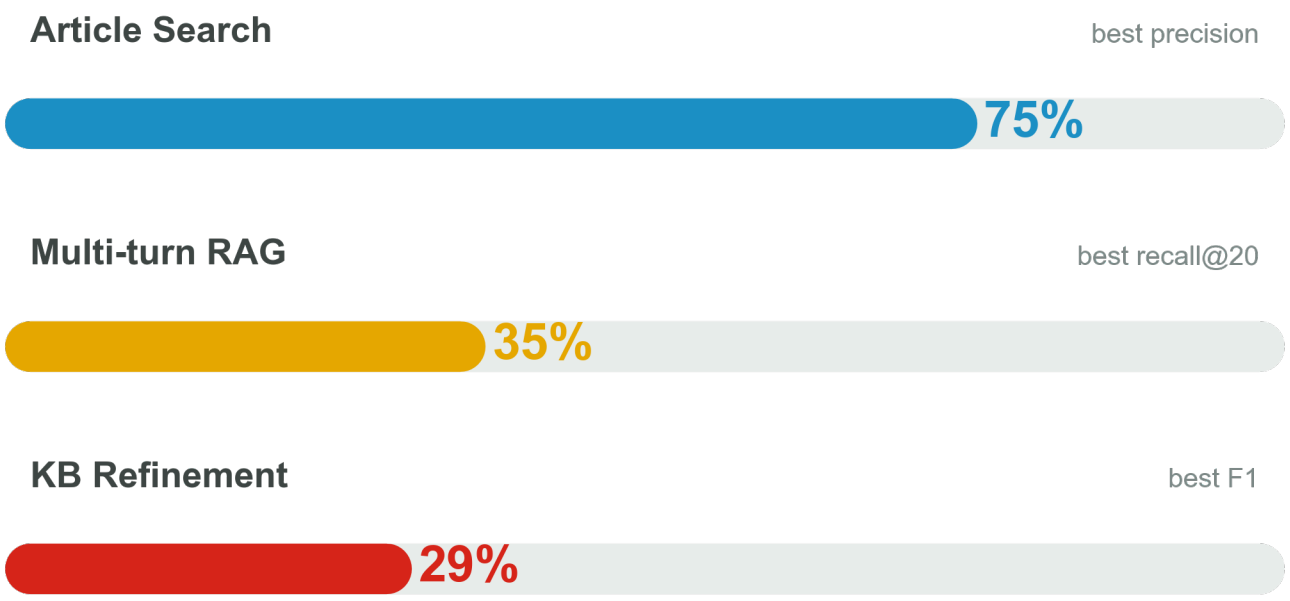
flagged similar or contradictory

Benchmark tasks

5

one per core contact-center operation

Even the best pipelines fall short on knowledge tasks



Best result achieved by any evaluated pipeline · bar track = 100%

READING THE GAP

68–75%

ceiling on closed-KB search — off-the-shelf retrieval isn't enough.

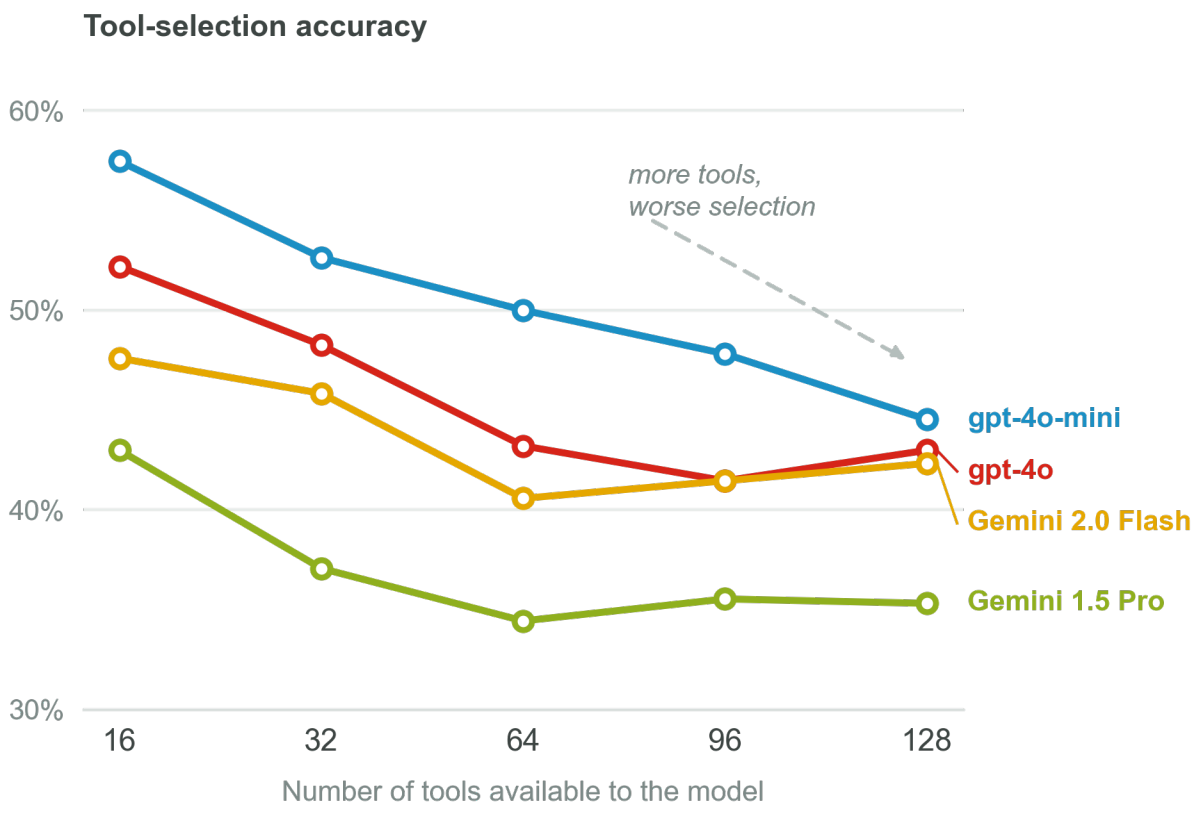
F1 0.29

on KB refinement: 0.85 precision but only 0.18 recall.

0.35

best recall@20 for context-aware multi-turn RAG.

Analytics is hard; tool selection degrades at scale



~30%

Intent accuracy on fine-grained Taxonomy A (208-class Tax B hits 71%).

32%

Top case-level Agent Quality — every check right in one call.

−13 pts

gpt-4o-mini's fall from 16 to 128 available tools.

TAKEAWAYS

What the numbers tell us

01

Off-the-shelf isn't enough

Conventional embeddings and single-shot prompting fail these tasks — real CXM needs purpose-built pipelines.

02

Grounding is the hard part

High hallucination and low recall show models struggle to stay strictly inside a domain KB.

03

Best value: Gemini 2.0 Flash

Strongest accuracy-per-dollar on analytical tasks; OpenAI embeddings still lead on retrieval.

BEYOND THE PAPER

Now multilingual and multi-domain — same pipeline, new flag

EN CXM_Arena

English

GreenBuild — home-improvement retail

3,168 KB articles

208 intents, largest taxonomy

979 intent-tagged conversations

FR CXM_Arena_French NEW

French

TerraBloom — luxury skincare & wellness

4,658 KB articles

286 intents in taxonomy

997 intent-tagged conversations

DE CXM_Arena_German NEW

German

Momentum — consumer electronics

4,277 KB articles

228 intents in taxonomy

499 intent-tagged conversations

Natively generated — not translated. A language flag re-runs the whole pipeline, so each suite gets its own brand, KBs, conversations, and labels from scratch — public on HuggingFace with baseline results and confidence intervals.

Further verticals already in the pipeline: gaming (Pixelbloom) · B2B SaaS (StreamlinePro).

SCOPE

Honest limitations

🧪 Synthetic by design

Despite injected noise, it can't fully capture the unpredictability of real customers; data inherits the generator LLMs' biases.

🌍 Coverage still growing

Three languages and three verticals so far — findings may not transfer to every industry or language yet.

🤖 Room for more baselines

We cover popular proprietary and open models, but more pipelines and models could be compared.

🚀 Where we're headed

Each new vertical or language is now just a pipeline config — more suites and community baselines next.

sprinklr

CONCLUSION

A benchmark for what CXM AI actually has to do.

- 5 operational tasks prior benchmarks ignored
- Realistic, noise-injected, KB-grounded — privacy-free
- Best models still fall short → open research runway
- Now in English, French & German — across three verticals

Get the datasets

huggingface.co/datasets/
sprinklr-huggingface/CXM_Arena
+ CXM_Arena_French · _German

Thank you · Kapil Sharma, Raghav Garg, Karan Gupta · Sprinklr · kapil.sharma@sprinklr.com

Backup slides

Full result tables & pipeline detail — for Q&A

BACKUP

KB Refinement — similarity detection (Table 2)

Embedding model	Precision	Recall	F1
text-embedding-3-large	**0.85**	**0.18**	**0.29**
jina-embeddings-v3	0.13	0.23	0.17
text-embedding-ada-002	0.10	0.31	0.15
all-mpnet-base-v2	0.16	0.13	0.15
multilingual-e5-large	0.06	0.47	0.11

High precision but very low recall → best F1 only 0.29. Contradiction detection needs non-embedding methods.

BACKUP

Intent Prediction — accuracy by taxonomy (Tables 3–4)

Model	Tax A (95)	Tax B (208)	Tax C (37)
GPT-4.1	**30.13**	**70.89**	**46.88**
Gemini 2.0 Flash	30.23	61.90	45.45
GPT-4.1 mini	29.11	62.21	44.84
Qwen-2.5-14B	25.64	49.95	34.83
GPT-4.1 nano	15.22	17.57	30.44
Qwen-2.5-7B	20.53	25.84	28.70

> Taxonomy B (208 intents, 2 levels) scores highest despite being largest — taxonomy quality matters more than size.

BACKUP

Agent Quality Adherence — accuracy (Table 5)

Model	Case-level (%)	Question-level (%)
GPT-4.1	**32.4**	**86.1**
GPT-4.1 mini	27.3	83.1
Gemini 2.0 Flash	26.3	83.7
GPT-4.1 nano	22.6	78.4
Qwen-2.5-14B	22.5	82.5
Qwen-2.5-7B	15.1	77.8

> Question-level looks healthy (~86%), but case-level — every check in a call correct — tops out at 32%.

BACKUP

Article Search & Multi-turn RAG retrieval (Tables 6, 8)

Article Search pipeline	Chunk	Precision
text-embedding-3-large	**1000**	**0.75**
ada-002 + BM25	500	0.68
text-embedding-ada-002	1000	0.67
e5-large-instruct	500	0.65
BM25	500	0.42

Multi-turn RAG (recall)	k	Recall
text-embedding-ada-002	**20**	**0.35**
text-embedding-3-large	20	0.34
ada-002	10	0.26
multilingual-e5-large	20	0.25
multilingual-e5-large	10	0.18

Retrieval caps at 0.75 precision; context-aware RAG recall@20 only ~0.35. Smaller chunks help slightly.

BACKUP

Multi-turn RAG — generation & tool calling (Tables 7, 9)

GENERATION QUALITY (%)

Generator	Correct	Incorrect	Halluc.	Refusal
Gemini 2.0 Flash	**61.0**	**5.0**	**13.0**	**21.0**
GPT-4o	50.0	4.0	24.0	22.0

TOOL-CALLING ACCURACY (%) BY NUMBER OF TOOLS

Model	16	32	64	96	128
gpt-4o-mini	**57.5**	**52.6**	**50.0**	**47.8**	**44.5**
gpt-4o	52.2	48.3	43.2	41.5	43.0
Gemini 2.0 Flash	47.6	45.8	40.6	41.5	42.3
Gemini 1.5 Pro	43.0	37.1	34.4	35.5	35.3

Gemini 2.0 Flash is the most grounded generator (13% hallucination). Tool accuracy falls as the tool set grows.