TECH_183: Generative AI for Mainframe: Real-World LLM Benchmarking, Cost Management, and Practical Results

Tim Ceradsky – Director Solution Engineering
Nick Guillemette – Solution Engineer

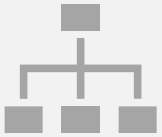
Copyright© by SHARE Association Except where otherwise noted, this work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 license. http://creativecommons.org/licenses/by-nc-nd/3.0/

From hype to hardware: benchmarking, cost, and real-world deployment

Generative AI for Mainframe

Why This Conversation Matters Now

GenAI has moved from experimentation into production discussions



Mainframe teams face higher risk, higher cost, and stricter governance

The question is no longer 'can we use AI?' but 'how do we use it safely?'

START WITH USE CASES, NOT MODELS

Viability comes before technology choices

Not All GenAI Use Cases Are Equal

- Explain and summarize workloads tolerate mistakes
- Chat assistants require fast response and trust
- Code and JCL generation demand near-zero error tolerance
- Some workflows should never be automated

Cost vs Criticality

Saving money is meaningless if output is unusable

Critical workflows prioritize accuracy and predictability

Non-critical tasks can optimize for lower cost

Engineering judgment must drive these decisions

SHARE
EDUCATE • NETWORK • INFLUENCE

Why traditional AI benchmarks fall short

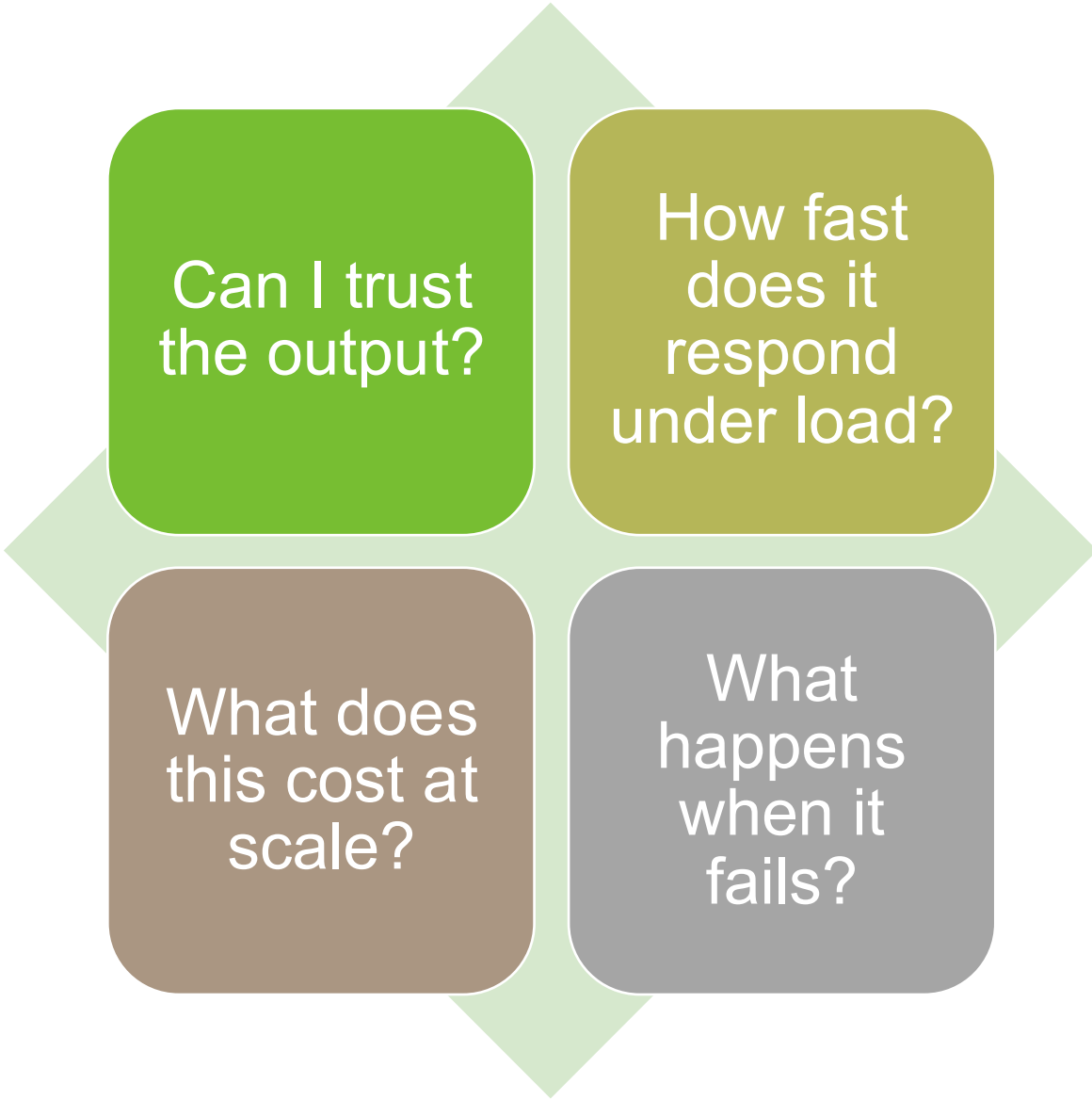
BENCHMARK LIKE AN ENGINEER

Copyright© by SHARE Association Except where otherwise noted, this work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 license. http://creativecommons.org/licenses/by-nc-nd/3.0/

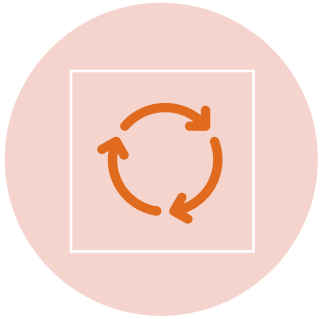
Why Traditional AI Benchmarks Fail Mainframe Teams

- Metrics like BLEU or MMLU measure language skill, not usefulness
- Accuracy alone ignores developer trust
- Latency is invisible in most AI benchmarks
- Production impact is never measured

What Actually Matters to Engineers

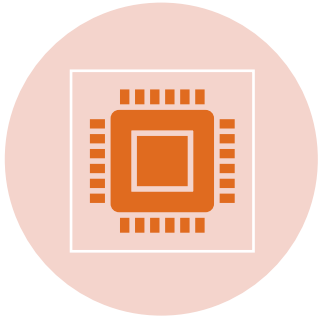


A Practical Benchmarking Framework



Task relevance: real COBOL, JCL, and domain-specific inputs

Quality scoring with human review



Operational metrics: latency, throughput, cost

Risk metrics: hallucinations and policy violations

Copyright© by SHARE Inc. Except where otherwise noted, this work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 license. http://creativecommons.org/licenses/by-nc-nd/3.0/

	query	reference_answer	rouge_recall_score	answer_correctness	answer_correctness_reasoning	faithfullness
2	what in devx code debug help me in cobol development?	DevX Code Debug greatly aids COB(	0.324817518	0.9	The generated answer effectively conveys the mai	1
3	How does BMC AMI Ops Insight utilize machine learning?	**BMC AMI Ops Insight utilizes ma(	0.307692308	0.9	The generated answer effectively conveys the mai	1
4	How to install BMC AMI Ops Monitor for Java Environments	To **install BMC AMI Ops Monitor f	0.4	0.8	The generated answer covers the main idea of ins	1
5	whats MJEINI00 do for ami ops monitor for java	The **MJEINI00** member is the m	0.448275862	0.8	The generated answer conveys the main idea that	1
6	What is Code insights	**BMC AMI DevX Code Insights** is	0.138157895	0.6	The generated answer captures the essence of Co	1
7	what are the General Guidelines for PARMLIB Members in c	The **general guidelines for PARMl	0.145695364	0.6	The generated answer provides some relevant gui	1
8	how to Initiate a primary command from the COMMAND lir	To **initiate a primary command fr	0.517006803	0.8	The generated answer provides a general method	0.75
9	How to Save and view historical data in BMC AMI OPs Moni	To **save and view historical data'	0.363157895	0.4	The generated answer provides some steps relate	0.5
10	How do I promote a task in code pipeline?	To **promote a task** in Code Pipe	0.436708861	0.8	The generated answer provides a comprehensive	1
11	What are the names of all the Code Pipeline started tasks?	### Code Pipeline Started Tasks**1	0.232394366	0.8	The generated answer accurately lists the main C(	1
12	In Code Pipeline how can I view task event history?	To view the **task event history** i	0.697478992	1	The generated answer accurately conveys the mai	1
13	How can I install DevX code pipeline?	To **install DevX Code Pipeline**, y	0.233516484	0.8	The generated answer covers the main ideas and	1
14	Using code insights program structure view how can i mak	To make a node the **root** in the	0.722891566	0.8	The generated answer conveys the main idea of m	1
15	what is required to install code insights	To install **BMC AMI DevX Code Ins	0.36196319	0.6	The generated answer provides some relevant inf(	1
16	What does "generate with parms" mean in code pipeline?	In **Code Pipeline**, "generate witl	0.321266968	0.8	The generated answer captures the main idea of '	1
17	what command is used to delete a task from a container ir	To delete a task from a container ir	0.620689655	0.8	The generated answer conveys the main idea of d	1
18	How can I use AI to explain a program using Code Insights	To use **AI to explain a program**	0.375	0.9	The generated answer effectively conveys the mai	1
19	How can I comment dead code using Code Insights?	To **comment out dead code** (as	0.545454545	0.9	The generated answer effectively conveys the mai	1
20	What is the logic flow view in code insights?	The **logic flow view** in Code Ins	0.457711443	0.8	The generated answer captures the main idea of t	1
21	How can I determine which programs I should refactor?	To determine which programs you s	0.388535032	0.9	The generated answer effectively conveys the mai	1
22	What is BMC AMI DevX Code Pipeline used for?	**BMC AMI DevX Code Pipeline** is	0.308219178	0.6	The generated answer captures some aspects of t	1
23	how to Initiate a primary command from the COMMAND lir	To **initiate a primary command fr	0.517006803	0.8	The generated answer provides a general method	0.75
24	How to Save and view historical data in BMC AMI OPs Moni	To **save and view historical data'	0.363157895	0.4	The generated answer provides some steps relate	0.5

Copyright© by SHARE Association Except where otherwise noted, this work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 license. http://creativecommons.org/licenses/by-nc-nd/3.0/

COST IS A FIRST-CLASS DESIGN CONSTRAINT

Why mainframe teams care more than anyone else

Bigger Models Rarely Mean Better Outcomes

- Large models increase latency and cost dramatically
- Smaller models often outperform on narrow tasks
- Right-sizing models improves reliability
- Prompt quality beats model tuning

Hardware & Model Tradeoffs

Model size drives GPU memory requirements

Service scalability vs LLM scalability

- Vertical vs Horizontal

CPU inference works for low-throughput workloads

Concurrency Changes Everything

- Latency degrades faster from concurrency than model size
- Batch and interactive workloads behave very differently
- GPU sharing enables multiple use cases on the same hardware
- Capacity planning must assume peak usage

DEPLOYMENT CHOICES

Cloud, on-prem, and hybrid realities

Cloud vs On-Prem vs Hybrid

- Cloud enables fast experimentation
- On-prem offers predictable long-term cost
- Hybrid allows gradual transition
- Adoption stage should drive deployment choice

Stages of GenAI Adoption

STAGE 1:
EXPERIMENTATION
IN THE CLOUD

STAGE 2:
BENCHMARKED
PILOTS

STAGE 3:
PLATFORM
DECISIONS

STAGE 4:
OPERATIONAL
GOVERNANCE


MODEL
STRATEGY
Public vs private models

Public Frontier Models

- ChatGPT, Claude, Gemini
- Best general-purpose capability
- Fast access and iteration
- Limited control and predictability

Private & On-Prem Models

Open-source models with local control

Data residency and IP protection

Predictable cost at scale

Requires operational maturity

Why Flexibility Matters

- Different workloads need different models
- Avoid vendor and model lock-in
- Hybrid strategies reduce risk
- Choice is the real advantage

THE INEVITABLE QUESTION

Why not just use ChatGPT?

Why Not Just Use ChatGPT?

COST GROWS QUICKLY AT SCALE

LIMITED CONTROL OVER DATA AND IP

LATENCY AND CONCURRENCY CONSTRAINTS

HARD TO INTEGRATE INTO GOVERNED WORKFLOWS

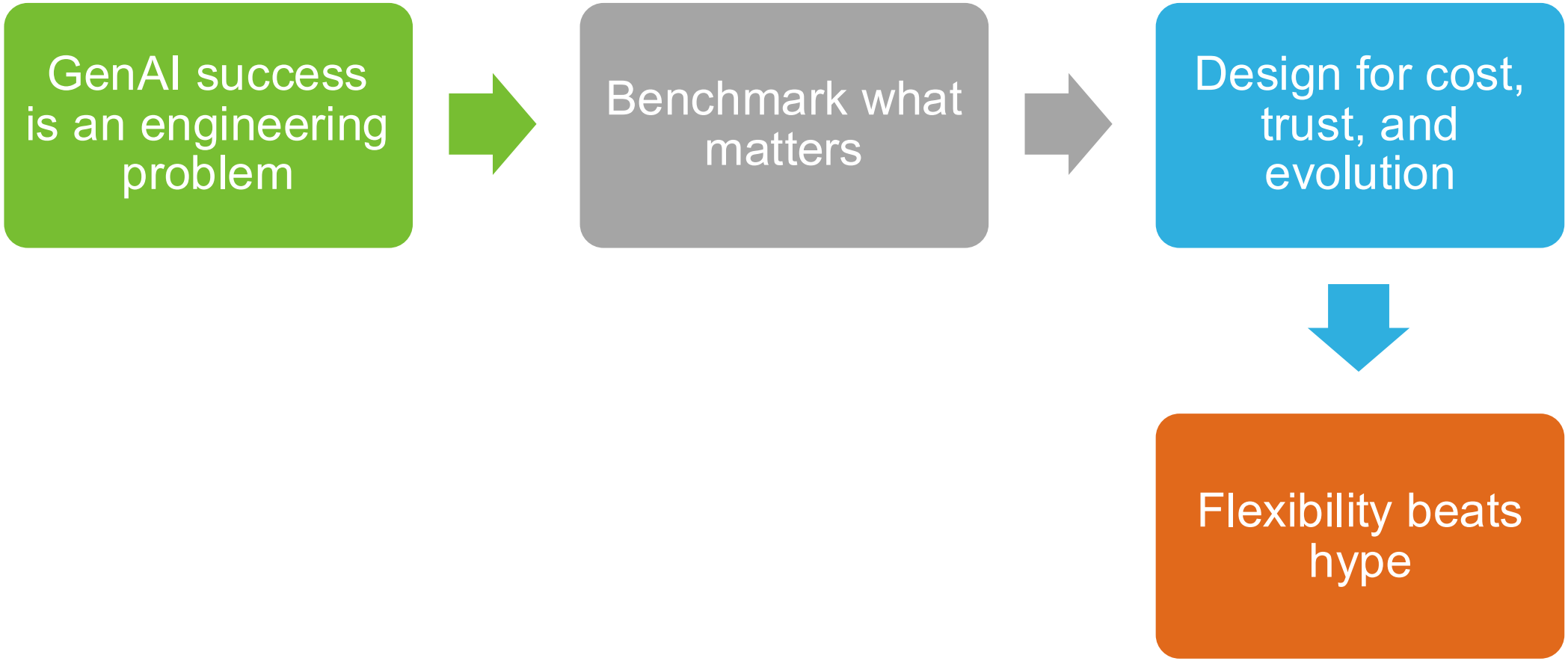
The Right Answer

ChatGPT is excellent for learning and experimentation

Enterprises outgrow it, not reject it

The goal is the right model for the right job

Final Takeaway



Contact

Connect on LinkedIn:

Tim Ceradsky

Tim_Ceradsky@bmc.com

Director – Solution Engineering

+1 618 978-7000

Your feedback is important!

SHARE
EDUCATE • NETWORK • INFLUENCE

Submit a session evaluation for each session you attend:

www.share.org/evaluation

Copyright© by SHARE Association Except where otherwise noted, this work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 license. http://creativecommons.org/licenses/by-nc-nd/3.0/