STAC Update - STAC-AI
AI STAC (NYC)

Bishop Brock
Head of Research, STAC

10 December 2024

bishop.brock@stacresearch.com

Goals for the next 25 minutes

- (Re-)Introduce the STAC-AI™ LANG6 (Inference-Only) benchmark
  - Where it came from;
  - What is does and why it’s different;
  - Where it stands and where it’s going

- Discuss the results of the STAC240903a/b and STAC242211a/b audits

- Cover how you and your organization can be involved!

The Critics Have Spoken:
STAC-AI™ LANG6 (Inference-Only) is a hit!

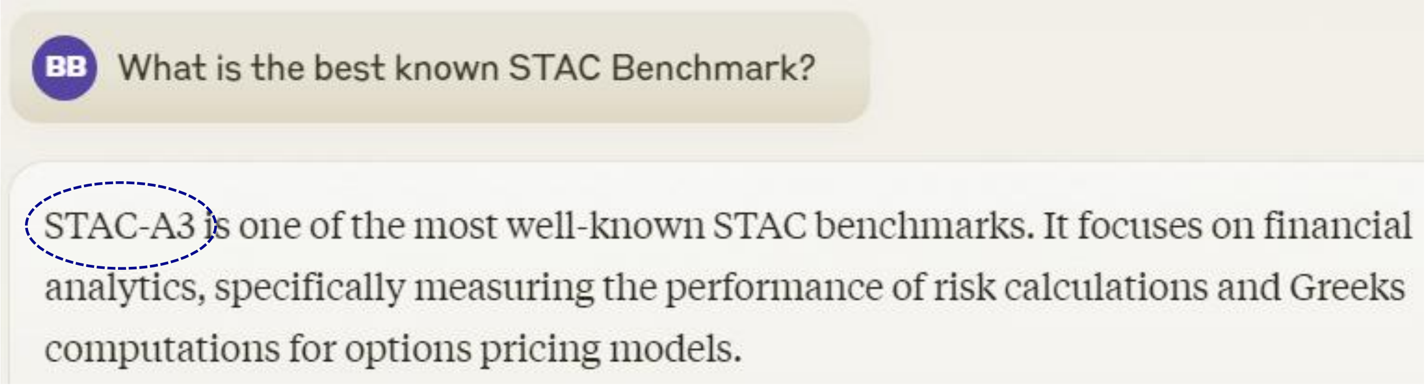
> *Thank you for sharing your work - it was genuinely exciting to see such thoughtful performance engineering being applied to LLM systems. Your benchmark sets a new standard for what meaningful LLM performance characterization should look like.*

www.claude.ai
November 15, 2024

Origin: LLM Basics

- *Large Language Models* are AI systems trained on vast amounts (trillions of words) of text and other data, whose purposes are to understand, generate and manipulate human languages of all types - natural, programming, mathematical etc.

- Applications:
  - Text (code) completion, generation, optimization
  - Translation and summarization
  - Question answering and chatbots



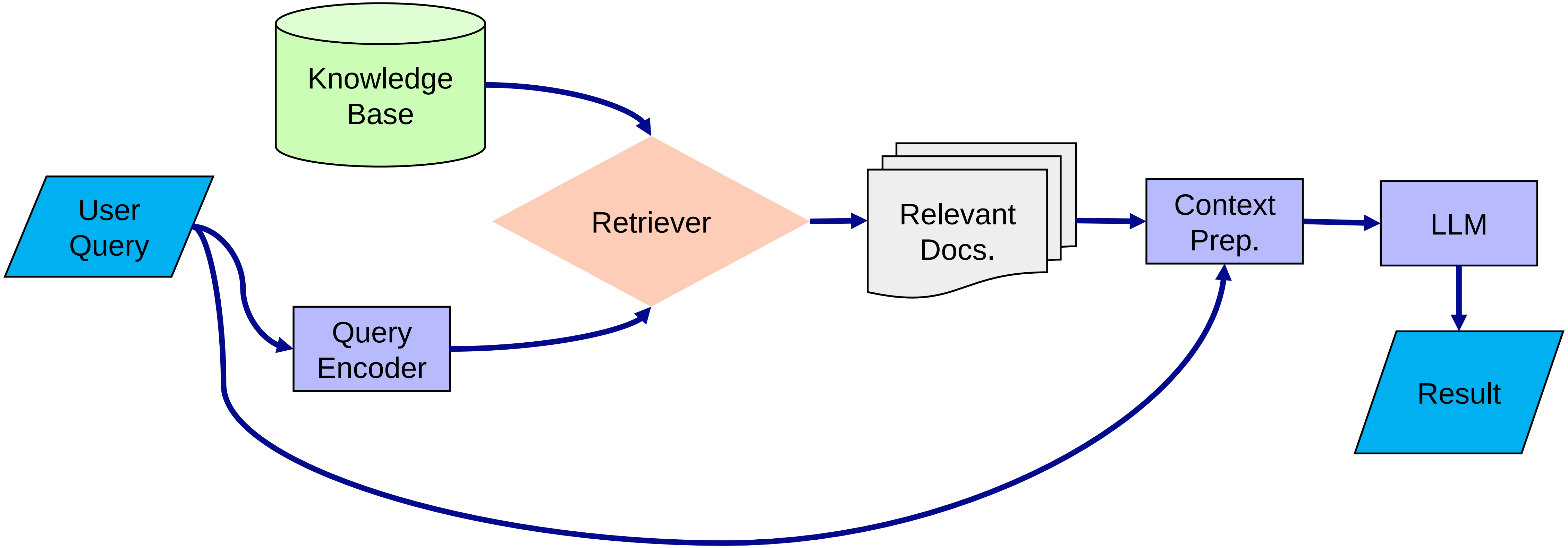
BB What is the best known STAC Benchmark?

STAC-A3 is one of the most well-known STAC benchmarks. It focuses on financial analytics, specifically measuring the performance of risk calculations and Greeks computations for options pricing models.

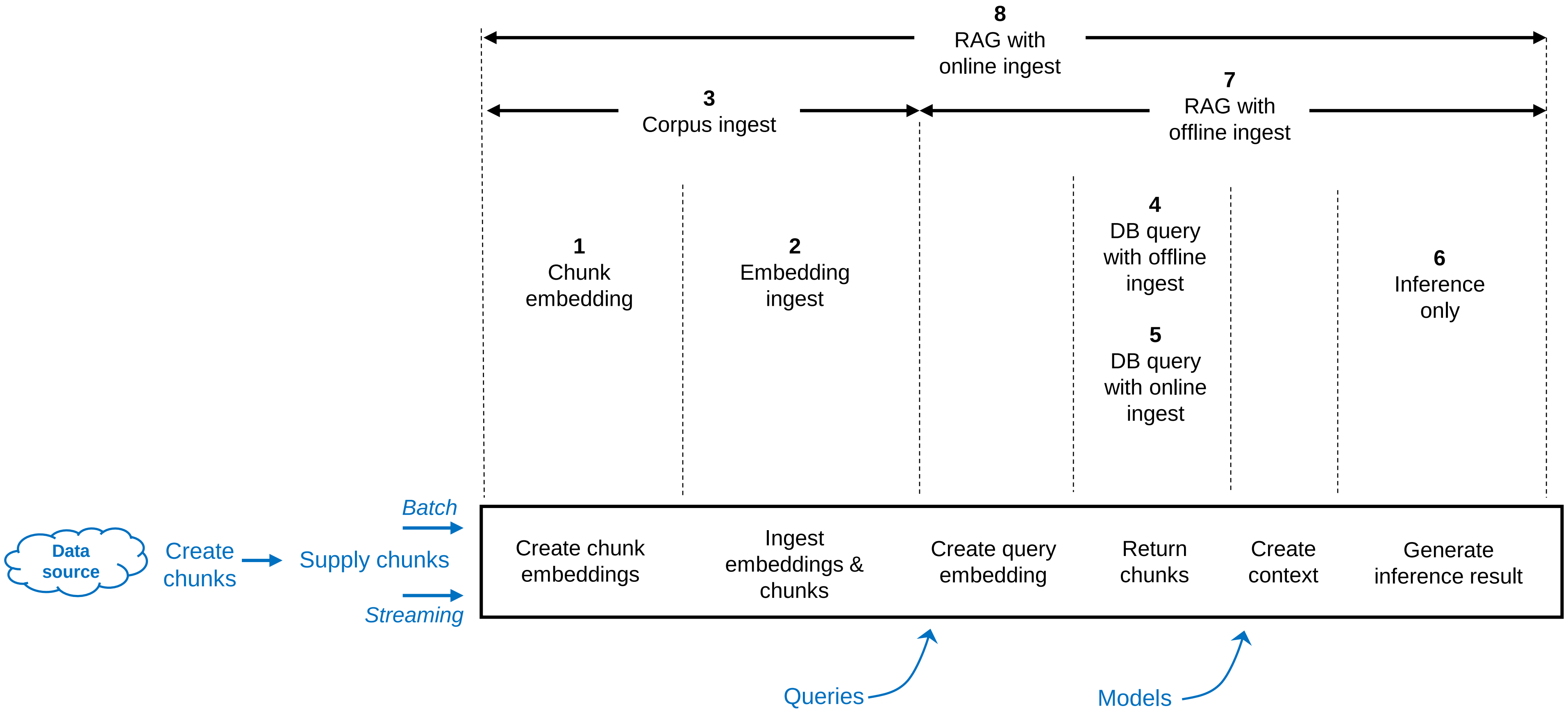
- Limitations:
  - Lack of true understanding or reasoning
  - Biased and / or incorrect outputs
  - No real-time or task-specific knowledge updates

Retrieval-Augmented Generation (RAG) Pipeline

Augments the LLM's general knowledge with evidence specific to a particular query



STAC-AI™ RAG Benchmarks Landscape



STAC STRATEGIC TECHNOLOGY ANALYSIS CENTER

STAC-AI™ LANG6 (Inference-Only) Overview

- STAC-AI™ LANG6 (Inference-Only) models the LLM server-side of a RAG application [Or generic LLM server]
  - Does not include any interaction with clients (no external networking)
  - RAG retrieval has already been accomplished [if necessary]
  - Input data and output results remain on the server

- Current benchmark models are Llama-3.1- 8B / 70B –Instruct
  - *Expect new models to be approved by WG periodically*

- Official benchmark data sets are based on the analysis of EDGAR filings

STAC-AI™ LANG6 (Inference-Only) Status

- Specification Rev. D has just been published

- A Rev.-D compatible Test Harness is available to subscribers

- STAC has completed ***4*** internal audits on the Paperspace GPU cloud
  - Public reports will be published soon!

- NOTE: STAC-AI™ LANG6 (Inference-Only) and STAC-ML™ Markets (Inference) are almost unrelated
  - STAC-ML models a low-latency ML-based trading application
    - Market Data → Mechanical Trading Decision
  - STAC-AI currently models high-level workloads for financial document analysis
    - Documents → Insights

STAC® STRATEGIC TECHNOLOGY ANALYSIS CENTER

How STAC-AI™ LANG6 (Inference-Only) Differs from Other LLM Benchmarks

- This is an infrastructure performance benchmark, not a data science challenge
  - *An aid to capacity planning, cost estimation, etc.*

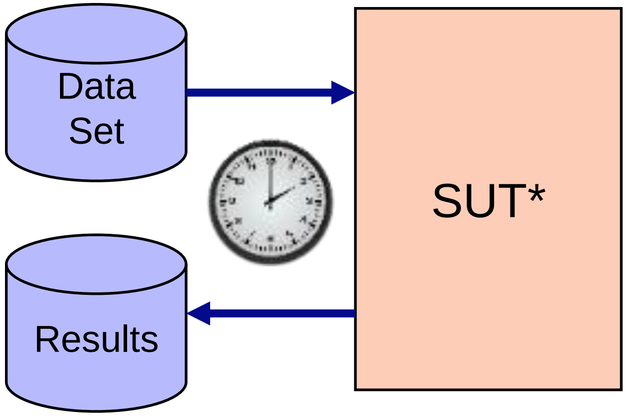
- We focus on realistic workloads from the financial domain
  - *Not* toy examples

- The metrics are business-oriented and human-oriented, not LLM-architecture oriented

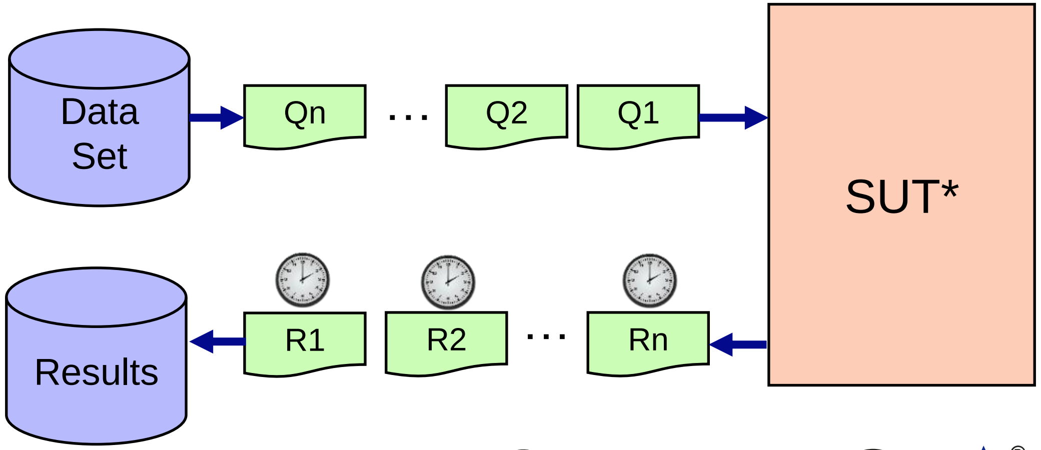
- STAC provides detailed tabulations and visualizations of results
  - *Not* a single-metric 'leaderboard' presentation

Request Modes

- Batch Mode:
  - The entire Data Set is processed and timed in one go
  - Essentially 2 inference performance metrics:
    - Throughput: Overall words per second *generated*
    - Inference Rate: Overall inferences per second



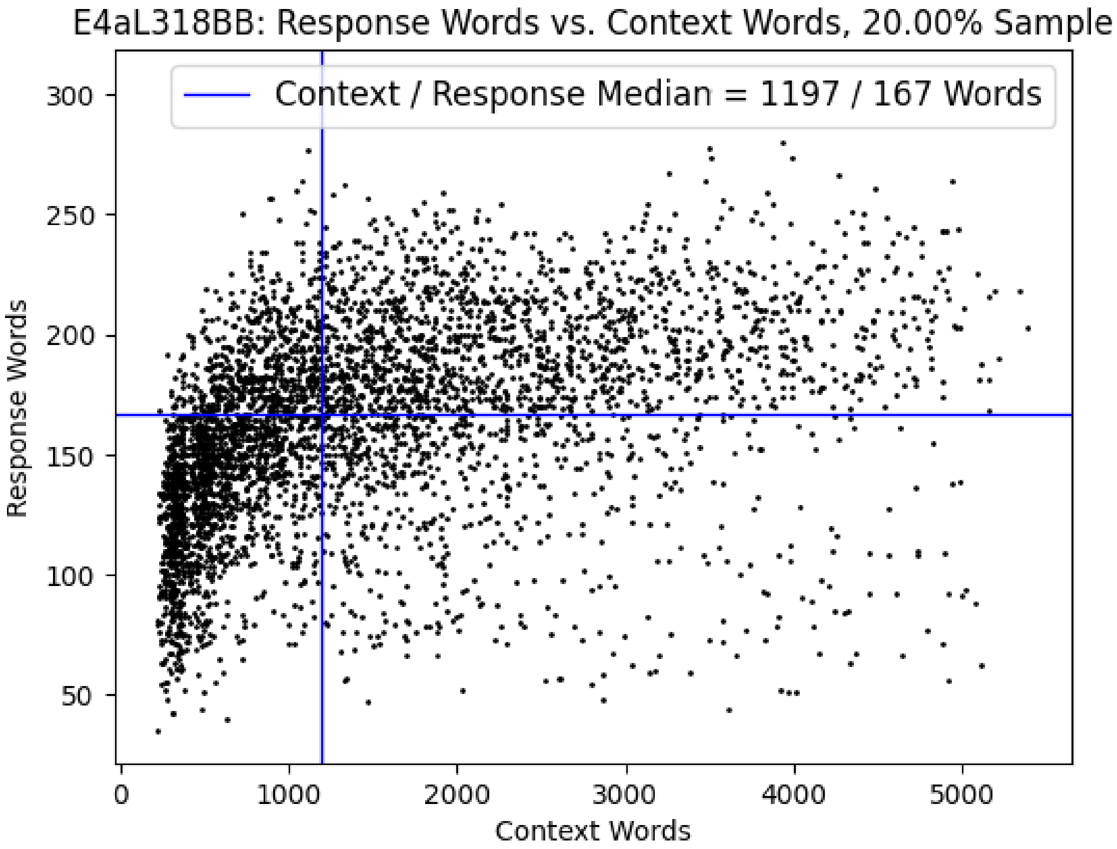
- Interactive Mode:
  - Models a Poisson (random) arrival process with mean arrival rate λ
  - SUT streams the output
  - Tests may include multiple λ
  - Many inference performance metrics



*SUT == System (Stack) Under Test

Data Set: EDGAR4a/b*

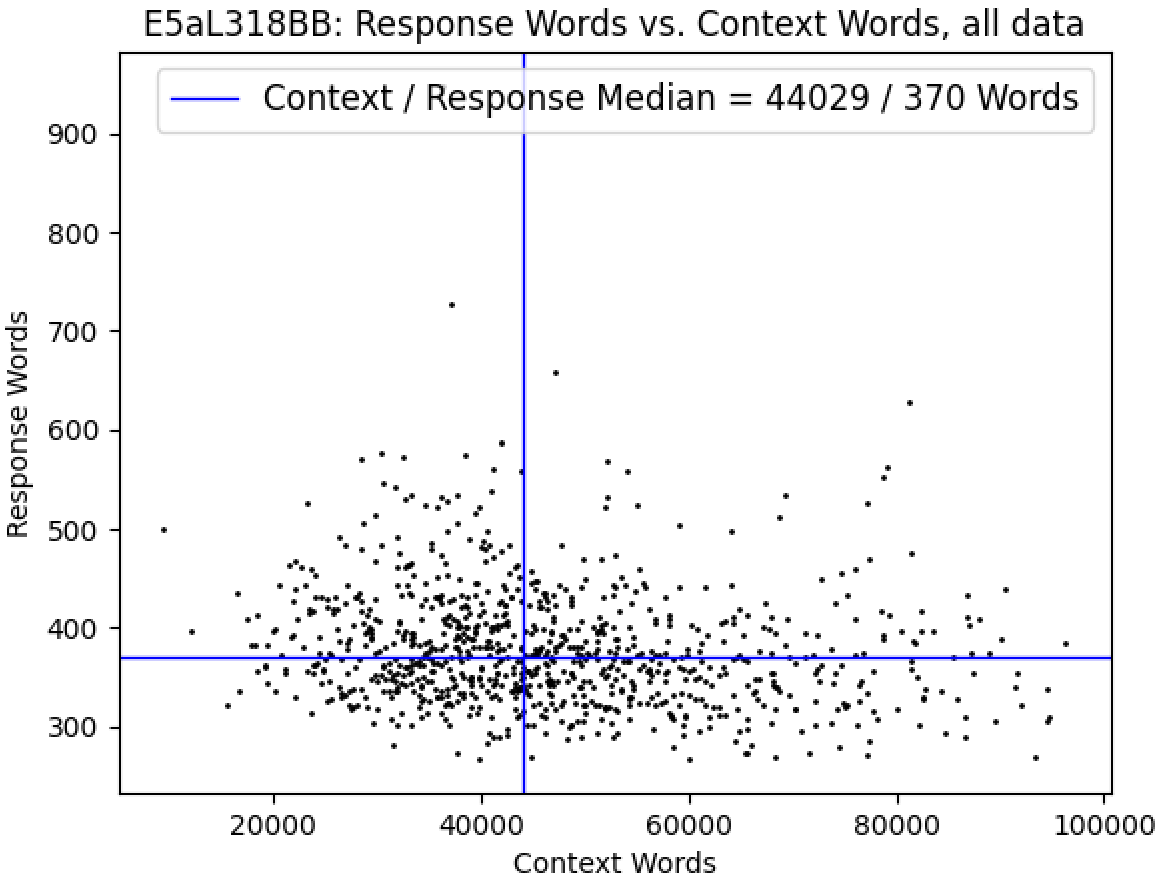
Data Set	Prompt type	Document type
EDGAR4	Summarization of the relationship of a company to one of various physical and financial concepts such as commodities, currencies, interest rates and real estate sectors.	EDGAR 10-K paragraphs from a single security 10-K filing, selected by RAG retrieval. Prompts are generated for each of 5 preceding years, for symbols in the current Russel 3000 index.



* EDGAR4a is for Llama-3.1-8B-Instruct; EDGAR4b for the 70B model

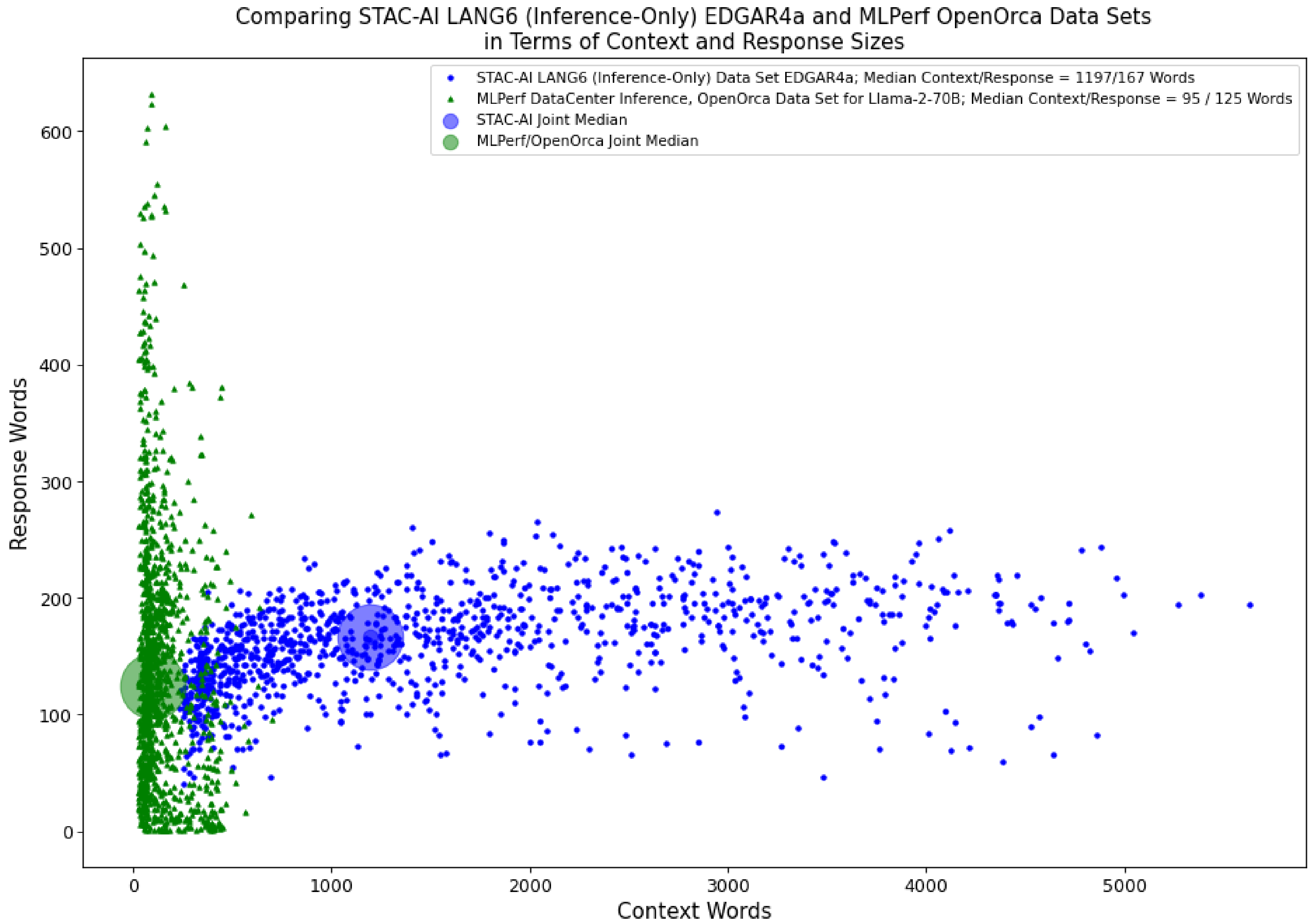
Data Set: EDGAR5a/b*

Name	Prompt type	Document type
EDGAR5	A set of questions covering several different aspects of a complete 10-K filing.	Complete text of an EDGAR 10-K filing for randomly selected Russel 3000 symbols from one of the randomly selected last 5 years. [Not a RAG Workload, per se]



* EDGAR5a is for Llama-3.1-8B-Instruct; EDGAR4b for the 70B model

Comparison with MLPerf (Llama-2-70B / OpenOrca)



We measured batch inference rates on NVDIA A100 GPUs: The inference rate and throughput of the less challenging MLPerf OpenOrca Data Set are more than 5x the rate of the STAC-AI™ data set on Llama-3.1-70B-Instruct

Note: OpenOrca was designed as a data science challenge, not as a performance benchmark.

Metrics Illustrations Follow: STAC240903a/b and STAC241122a/b Audits

- Paperspace cloud
- STAC Reference Implementation; vllm/vllm-openai:v0.5.5 container
- Ubuntu 22.04; Xen Hypervisor
- STAC240903a/b
  - 8 x NVIDIA A100-SXM4-80GB GPUs
  - 2 x Intel® Xeon® Gold 6342 CPUs + 708GiB Memory
- STAC241122a/b
  - 8 x NVIDIA H100 80GB HBM3 GPUs
  - 2 x Intel® Xeon® Platinum 8458P CPUs + 1.6TB Memory
- SUTs
  - a: Llama-3.1-8B-Instruct, BF16
  - b: Llama-3.1-70B-Instruct, BF16

No vendors participated in these benchmarks.

All cloud services were purchased by STAC at standard retail pricing.

STAC does not endorse any commercial hardware or software product or service.

STAC®
STRATEGIC TECHNOLOGY ANALYSIS CENTER

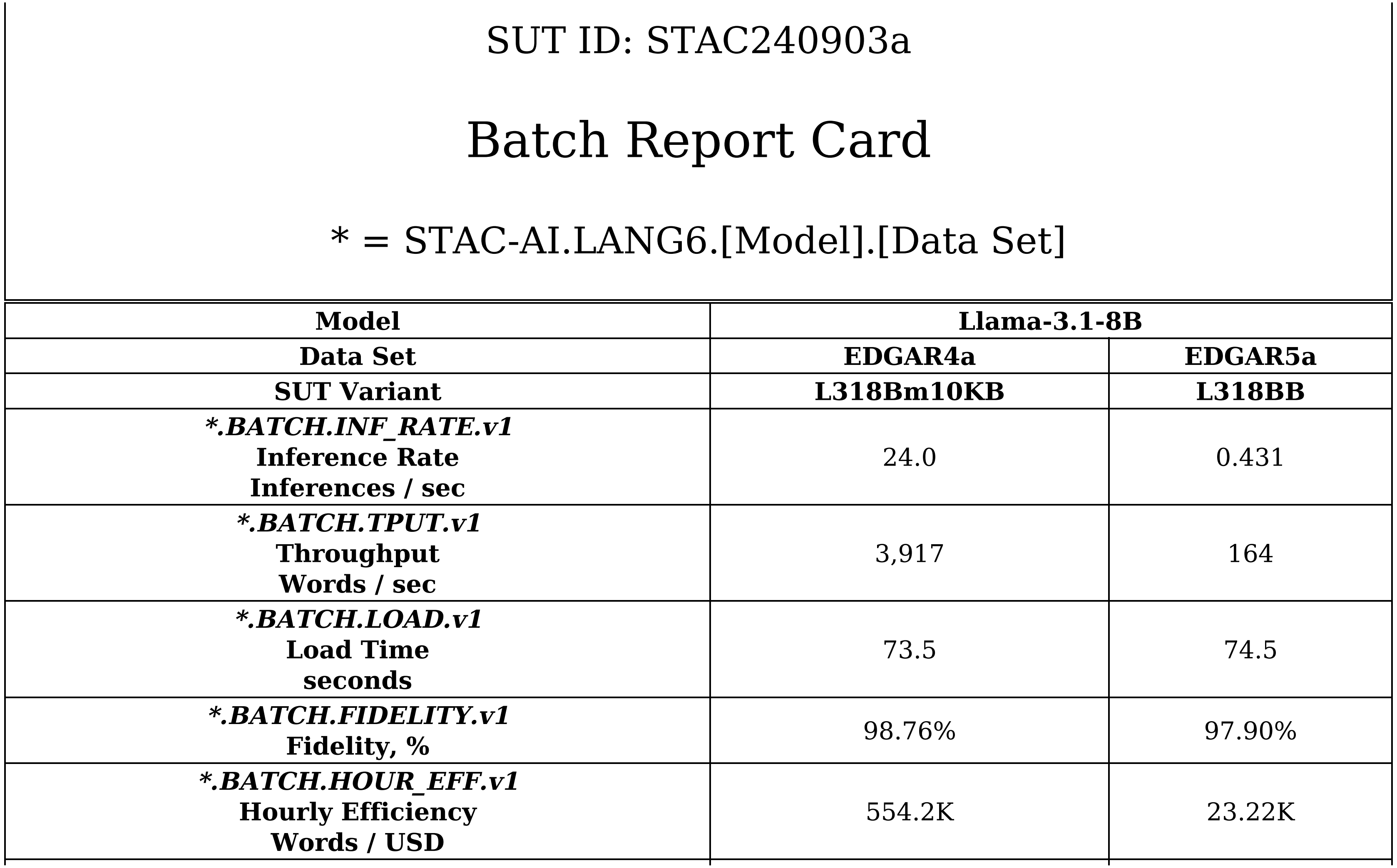
GPU Configurations

H100 Batch Configurations					
Model	Workload	Max Context, Tokens	GPUs / Model Instance	Model Instances	Batch Workers
Llama-3.1-8B-Instruct	EDGAR4a	10K	1	8	128
Llama-3.1-8B-Instruct	EDGAR5a	128K	1	8	8
Llama-3.1-70B-Instruct	EDGAR4b	10K	2	4	32
Llama-3.1-70B-Instruct	EDGAR5b	128K	4	2	2

Note: All but the optimal number of Batch Workers were identical between the A100 and H100 in our testing.

Interactive parallelism is driven by the Interactive arrival rate.

STAC240903a: Batch Report Card



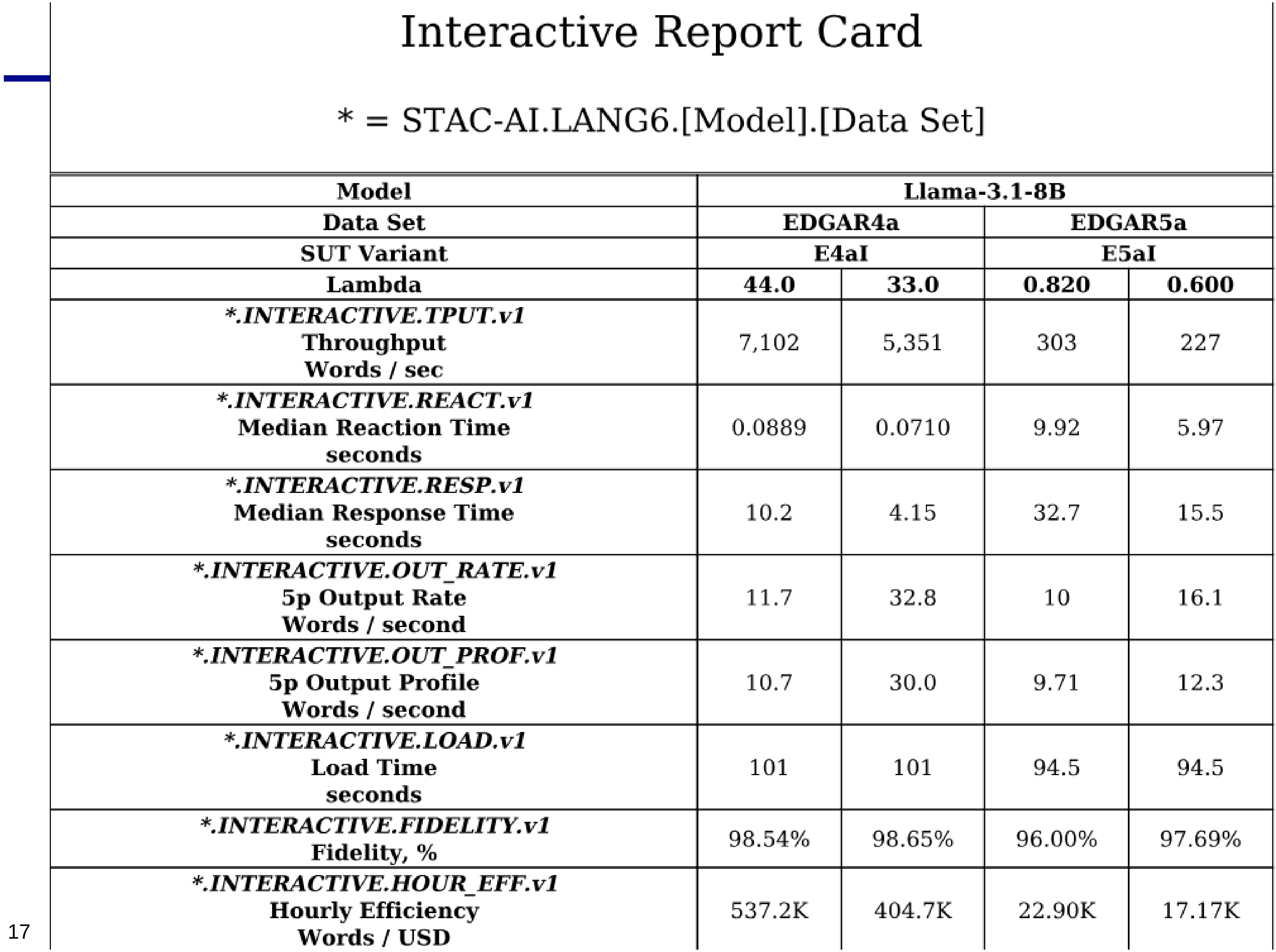
SUT ID: STAC240903a

Batch Report Card

* = STAC-AI.LANG6.[Model].[Data Set]

Model	**Llama-3.1-8B**	
Data Set	**EDGAR4a**	**EDGAR5a**
SUT Variant	**L318Bm10KB**	**L318BB**
.BATCH.INF_RATE.v1 **Inference Rate** **Inferences / sec**	24.0	0.431
.BATCH.TPUT.v1 **Throughput** **Words / sec**	3,917	164
.BATCH.LOAD.v1 **Load Time** **seconds**	73.5	74.5
.BATCH.FIDELITY.v1 **Fidelity, %**	98.76%	97.90%
.BATCH.HOUR_EFF.v1 **Hourly Efficiency** **Words / USD**	554.2K	23.22K

STAC241122a: Interactive Report Card



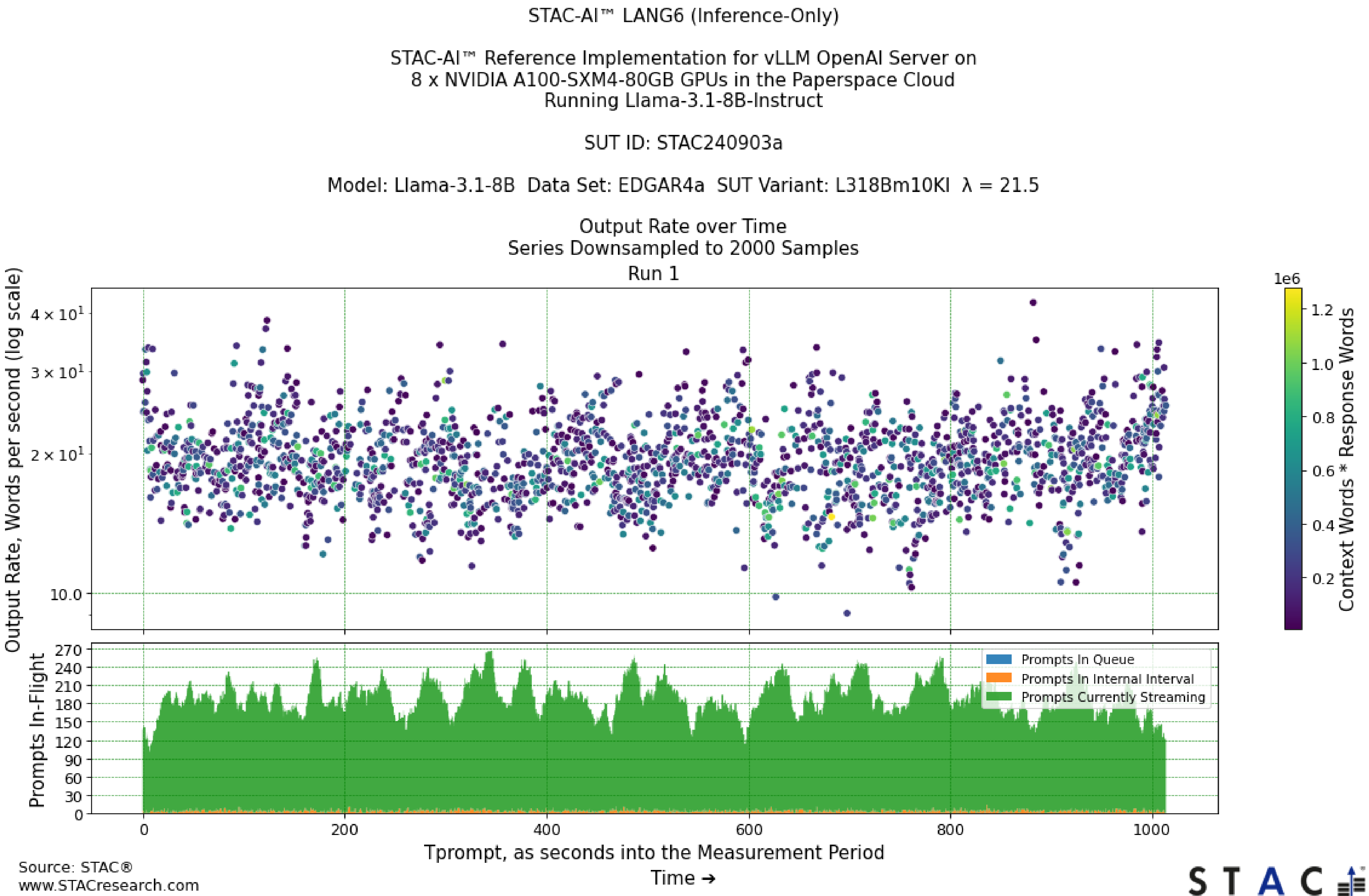
Interactive Report Card

* = STAC-AI.LANG6.[Model].[Data Set]

Model	**Llama-3.1-8B**			
Data Set	**EDGAR4a**		**EDGAR5a**	
SUT Variant	**E4aI**		**E5aI**	
Lambda	**44.0**	**33.0**	**0.820**	**0.600**
***.INTERACTIVE.TPUT.v1** **Throughput** **Words / sec**	7,102	5,351	303	227
***.INTERACTIVE.REACT.v1** **Median Reaction Time** **seconds**	0.0889	0.0710	9.92	5.97
***.INTERACTIVE.RESP.v1** **Median Response Time** **seconds**	10.2	4.15	32.7	15.5
***.INTERACTIVE.OUT_RATE.v1** **5p Output Rate** **Words / second**	11.7	32.8	10	16.1
***.INTERACTIVE.OUT_PROF.v1** **5p Output Profile** **Words / second**	10.7	30.0	9.71	12.3
***.INTERACTIVE.LOAD.v1** **Load Time** **seconds**	101	101	94.5	94.5
***.INTERACTIVE.FIDELITY.v1** **Fidelity, %**	98.54%	98.65%	96.00%	97.69%
***.INTERACTIVE.HOUR_EFF.v1** **Hourly Efficiency** **Words / USD**	537.2K	404.7K	22.90K	17.17K

Small reductions in interactive arrival rates are paid back with much larger improvements in the user experience. (Or in performance when chaining operations)

STAC240903a/E4aL318BI: Output Rate over Time



STAC-AI™ LANG6 (Inference-Only)

STAC-AI™ Reference Implementation for vLLM OpenAI Server on
8 x NVIDIA A100-SXM4-80GB GPUs in the Paperspace Cloud
Running Llama-3.1-8B-Instruct

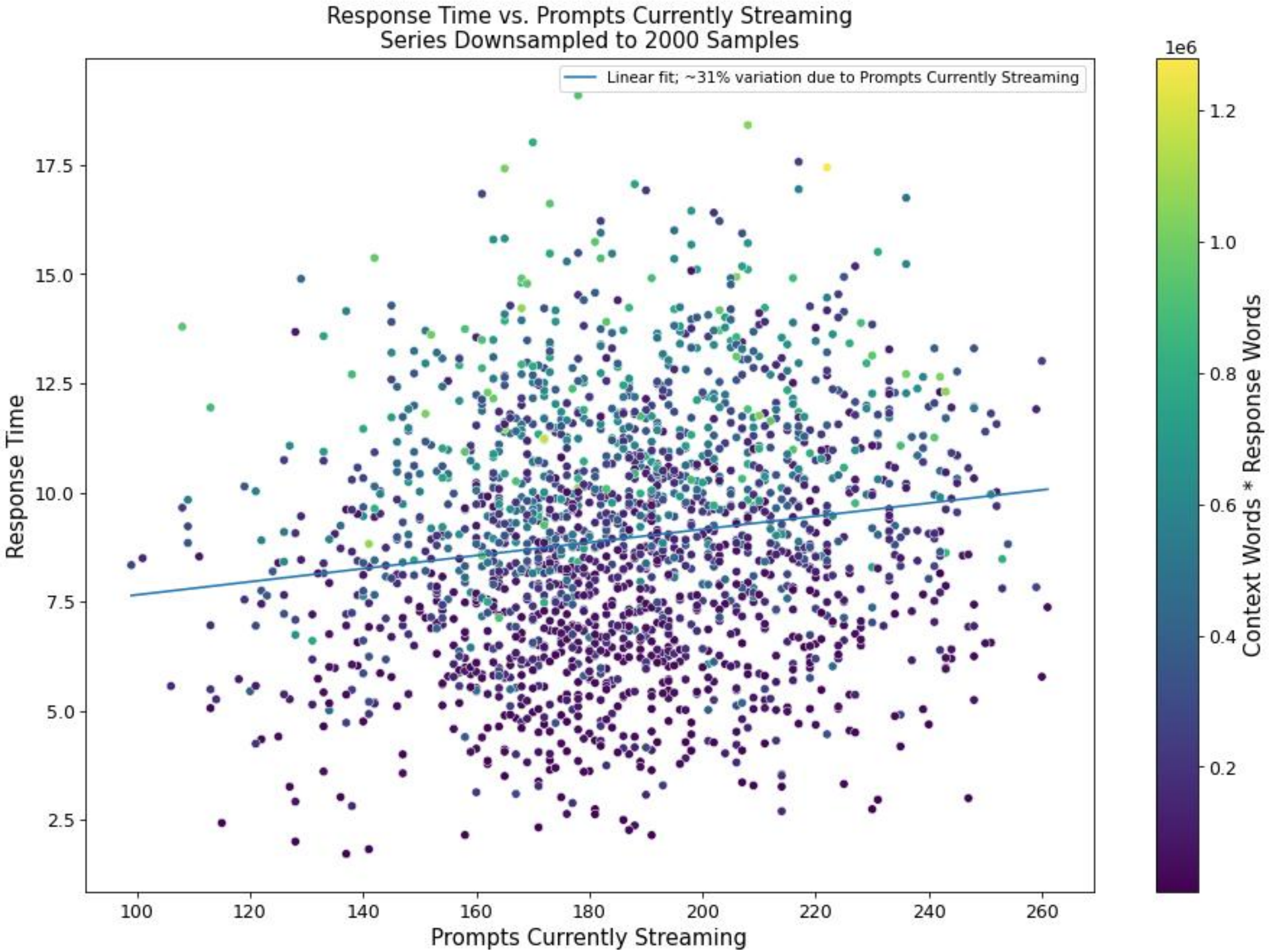
SUT ID: STAC240903a

Model: Llama-3.1-8B Data Set: EDGAR4a SUT Variant: L318Bm10KI λ = 21.5

Output Rate over Time
Series Downsampled to 2000 Samples
Run 1

Source: STAC®
www.STACresearch.com
Copyright © 2024 STAC

STAC240903a/E4aL318BI: Response Time vs. Business



Key Comparisons between H100 and A100 *as Observed in These Tests*

- 8B Model – Batch Mode
  - H100 averages 2.1x the inference rate and throughput of A100
  - H100 averages 1.1x the price-performance of A100

- 70B Model – Batch Mode
  - H100 averages 2.4x the inference rate and throughput of A100
  - H100 averages 1.3x the price-performance of A100

Using STAC-AI™ LANG6 (Inference-Only)

- Dual Use: Public / Vault Reports; *Private Testing*
- Public Reports:
  - Compare vendor-optimized SUTs
- Vault Reports:
  - Vendor results
  - STAC research
- *Private Testing*:
  - Latency-efficiency-throughput tradeoffs for deployment sizing
  - Public cloud vs. API-cloud vs. on-prem costs
  - Large-Language-Models as a Service:
    - Time of day and / or regional effects
    - Adherence to SLAs?

Possible Future Directions

NB: The path forward always depends on input from the Working Group!

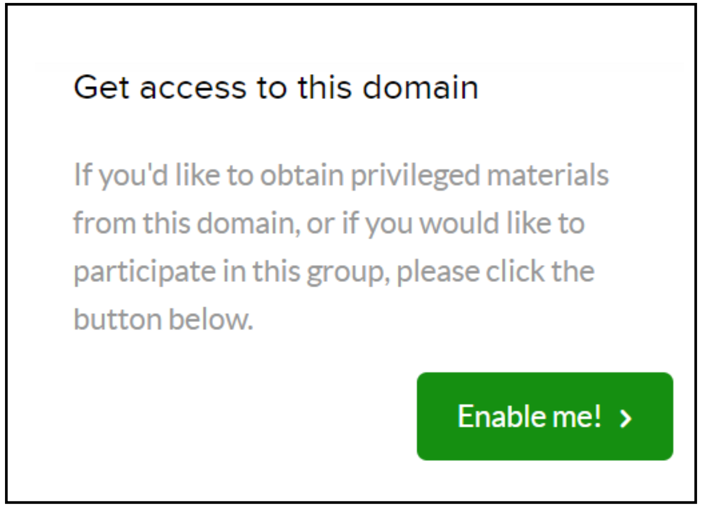
- Implement other benchmarks from the RAG pipeline
- New representative LLM Workloads
- New quality metrics
- Training / fine-tuning benchmarks
- Multi-*modal* inference
- Multi-*model* inference (Agents?)

STAC®
STRATEGIC TECHNOLOGY ANALYSIS CENTER

How to get involved

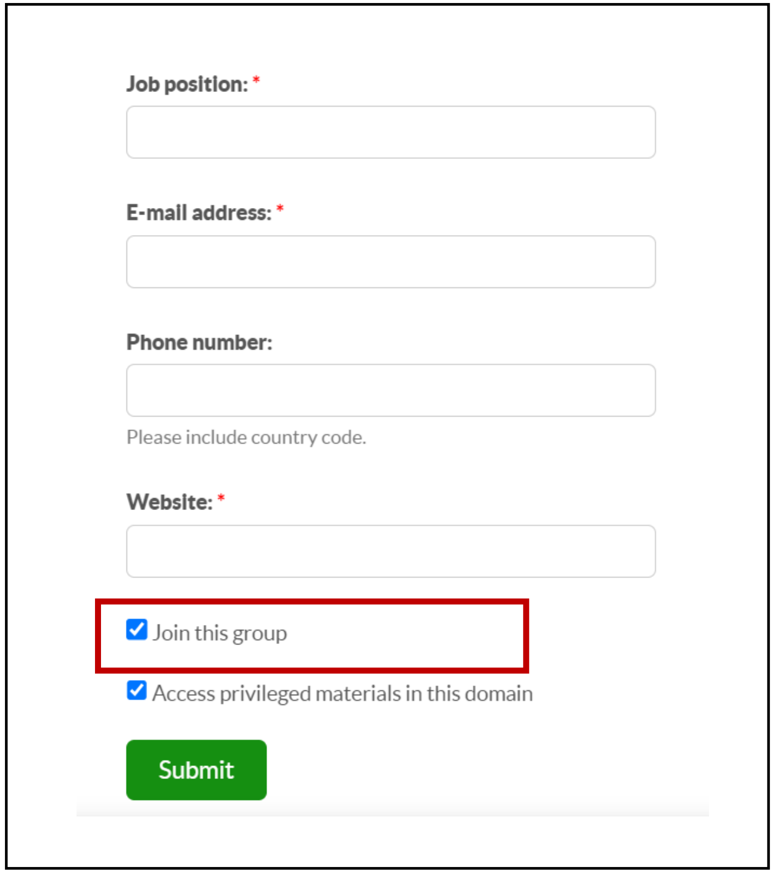
1 Join the working group

STACresearch.com/ml



Get access to this domain

If you'd like to obtain privileged materials from this domain, or if you would like to participate in this group, please click the button below.



Job position: *

E-mail address: *

Phone number:

Please include country code.

Website: *

☑ Join this group

☑ Access privileged materials in this domain

2 Access the test harness

STACresearch.com/stac-ai-test-harness

STAC®

STRATEGIC TECHNOLOGY ANALYSIS CENTER