L40S GPU Stress Test Results

Technical Analysis and Hardware Recommendations

Executive Summary

Current Hardware Cannot Support Production Deployment

Key Finding

L40S GPU is fundamentally inadequate for multi-user AI services due to critical performance limitations.

Critical Metrics

- 80% failure rate at 32 concurrent users.
- Response time degradation from 3.5 seconds to 33+ seconds.
- 91.3% VRAM saturation before any user requests.

Bottom Line

The L40S represents a technical dead-end for production deployment of concurrent AI services. Immediate hardware upgrade is required.

Research Methodology

Formal Hypothesis Testing Approach

Null Hypothesis (H_0)

Increasing concurrent inference requests on L40S GPU has no significant negative effect on system failure rate, response latency, or semantic quality of Llama 3.1 8B model outputs.

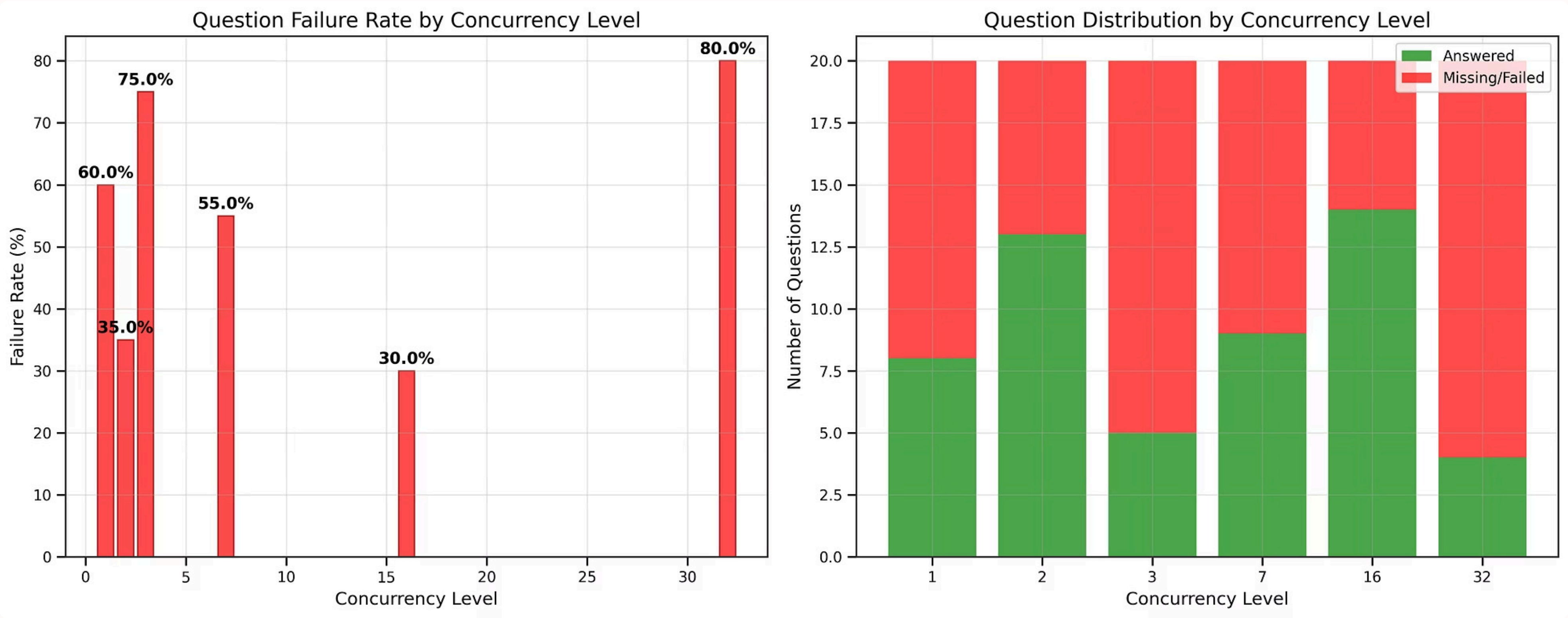
Alternative Hypothesis (H_1)

Increasing concurrent requests will cause significant increases in failure rates and response latency, plus measurable degradation in response quality, coherence, and accuracy, resulting in hallucinations and unusable output.

Test Design: Systematic stress testing was conducted from 1 to 32 concurrent users, meticulously collecting comprehensive performance and quality metrics to validate our hypotheses. This approach allowed us to identify critical bottlenecks and assess system behavior under realistic load conditions.

Finding 1 - System Reliability Collapse

Extreme Failure Rates Make Service Unusable



Question Failure Rate by Concurrency Level

Question Distribution by Concurrency Level

Under load, the L40S GPU exhibits **catastrophic reliability failures**, rendering the service functionally useless for multi-user scenarios. This is not a partial degradation but a complete breakdown.

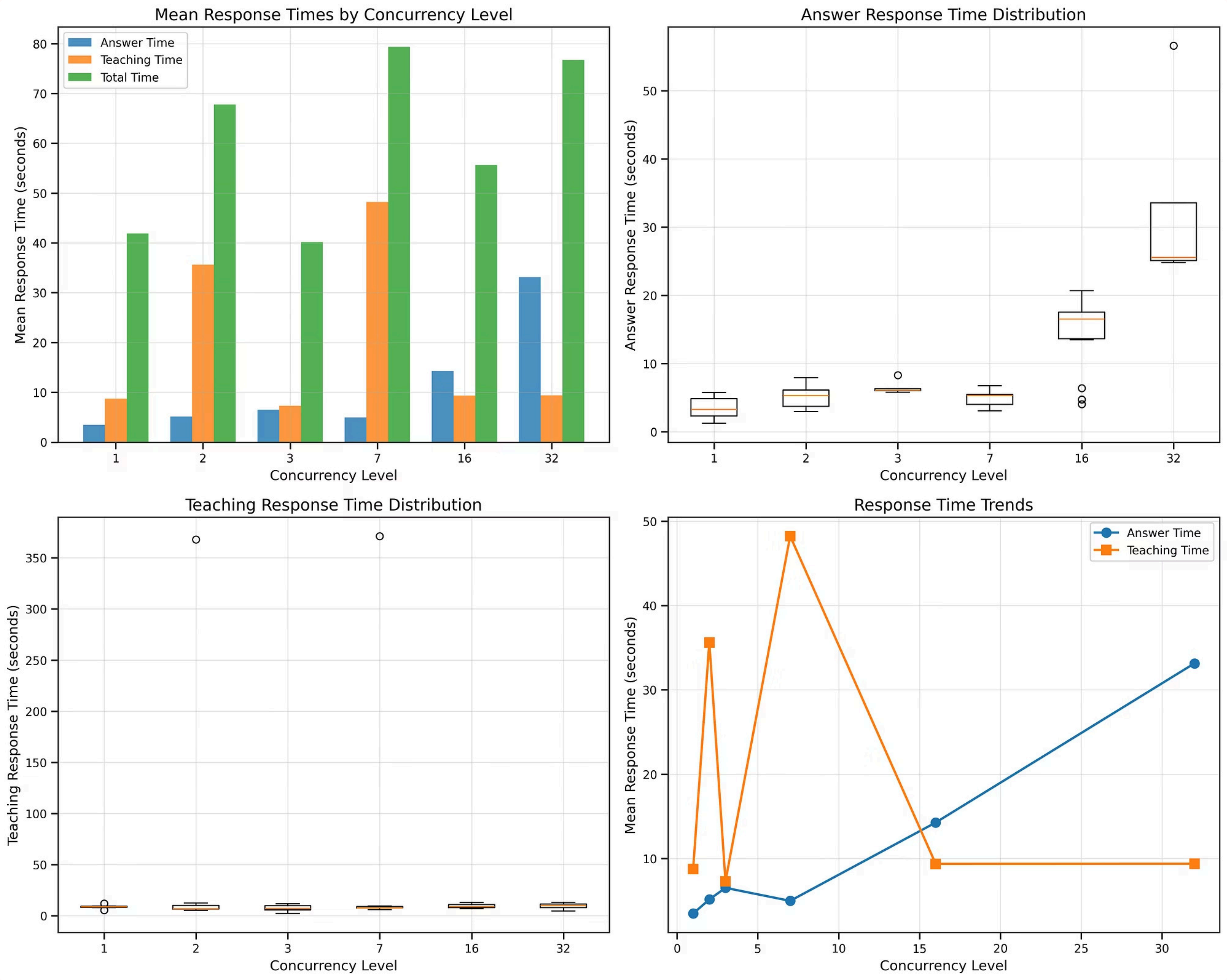
- **3 concurrent users:** 75% failure rate
- **32 concurrent users:** 80% failure rate
- **4 out of 5 user queries** receive no response at peak load.

Business Impact

- Cannot deploy a service that fails for the majority of users. This level of unreliability would immediately destroy user trust and render the application functionally useless.
- At high concurrency the llm was able to respond , but in flawed way , exposing internal prompt, changing language , repeating prompt or question.

Finding 2 - Performance Degradation

Response Times Become Unacceptable Under Load



- **Single user:** 3.5 seconds average response time
- **32 users:** 33+ seconds average response time
- **10x performance degradation** at scale, far beyond acceptable limits for interactive AI.

The performance degradation is **exponential**, not linear. Some users experience reasonable wait times, while others face delays of nearly a minute, leading to an inconsistent and frustrating experience.

⊗ Business Impact

- Slow and unpredictable service leads to user frustration and abandonment. An AI assistant slower than manual work defeats the purpose of automation and provides no tangible value.
- When load is too much the model does not answer , in the case of 32 concurrent requests the model gave incoherent answers.

Finding 3 - Quality Deterioration

AI Output Becomes Unreliable and Incoherent

Accuracy Volatility

Accuracy scores become highly volatile under load, with unpredictable dips and spikes, making outputs unreliable.

Context Loss

The model frequently loses the ability to maintain conversational context, leading to disjointed and irrelevant responses.

Response Incoherence

A systematic breakdown in response coherence is observed, resulting in nonsensical or contradictory outputs.

Observable Failure Modes: The model "forgets" conversation context mid-response, generates contradictory or nonsensical outputs, and produces factually incorrect information (hallucinations). This is not just a performance issue; it's a fundamental **quality collapse**.

Business Impact

- Unreliable AI output is worse than no AI – it actively misleads users and requires human verification, eliminating any potential efficiency gains and introducing new risks.

Quality Failure Analysis - Language Hallucination

AI Responds in Wrong Language Under Load

32 Concurrent requests English:

This occurs when the AI responds in a language different from the question's language.

- **question_id: 2**
  - **Evidence:** The question is in English, but the ai_reasoning is in Malay: "Pada notis, terdapat maklumat yang menyatakan bahawa..."
- **question_id: 15**
  - **Evidence:** The question is in English, but the ai_answer uses the Malay word for "Answer": "Jawapan: B"
- **question_id: 18**
  - **Evidence:** The question is in English, but the ai_reasoning is in Malay: "Pada soal ini, kita diminta untuk menganalisis keperluan untuk menghadapi soal dan jawapan yang mengandung."
- **question_id: 19**
  - **Evidence:** The question is in English, but the ai_answer and ai_reasoning are in Malay. The AI incorrectly states, "The context is in Malay, so I will respond entirely in Malay."

16 Concurrent requests History:

- **question_id: 6 (eval_type: Teaching)**
  - **Evidence:** The question is in Malay, but the ai_reasoning is in English: "The correct answer is A because the British introduced the Malayan Union..."
- **question_id: 8 (eval_type: Teaching)**
  - **Evidence:** The question is in Malay, but the ai_reasoning is in English: "The correct answer is B because it aligns with the context of the passage."
- **question_id: 9 (eval_type: Teaching)**
  - **Evidence:** The question is in Malay, but the ai_reasoning is in English: "The correct answer is D because the question asks about the leader of the Federation of Malaya in 1948."
- **question_id: 10 (eval_type: Teaching)**
  - **Evidence:** The question is in Malay, but the ai_reasoning is a mix of English and Malay: "The correct answer is B because Parti Komunis Malaya menghakis sokongan..."
- **question_id: 11 (eval_type: Teaching)**
  - **Evidence:** The question is in Malay, but the ai_reasoning is in English: "The correct answer is A because Rancangan Briggs successfully weakened the communist movement..."

Technical Cause: Severe memory pressure on the L40S GPU forces the model to lose language context awareness, leading to erratic linguistic behavior.

⊗ Business Impact

- Highly unprofessional for a corporate tool. Makes the system appear broken and untrustworthy, eroding user confidence immediately and reflecting poorly on the brand.

Quality Failure Analysis - Contradictory Logic

AI Provides Wrong Answer While Explaining Correct One

32 Concurrent requests English:

- **question_id: 5 (eval_type: Teaching)**
  - **Evidence:** The AI's chosen answer was **A**, but the ai_reasoning argues for the correct answer, **C**: "The correct answer is 6. C because the article states that..."
- **question_id: 8 (eval_type: Answer)**
  - **Evidence:** The ai_answer is **C**, but the ai_reasoning explains why **B** is correct: "This shows that he has to work every day... Therefore, option B is the correct answer."
- **question_id: 8 (eval_type: Teaching)**
  - **Evidence:** The AI's chosen answer was **C**, but the ai_reasoning is a detailed explanation for why **B** is correct: "Let's break down the question and the passage to understand why option B is the correct answer."
- **question_id: 15 (eval_type: Teaching)**
  - **Evidence:** The AI's chosen answer was **B**, but the ai_reasoning argues for the correct answer, **F**: "The correct answer is 33. F because it accurately reflects the idea that..."
- **question_id: 16 (eval_type: Teaching)**
  - **Evidence:** The AI's chosen answer was **A**, but the ai_reasoning explains why **D** is correct: "The correct answer is 34. D because it highlights the potential economic impact..."

16 Concurrent requests History:

- **question_id: 0**
  - **Evidence:** The AI claims the answer is B (Warfare) but then states in its reasoning that the text provides no evidence for how the territories were conquered: "...tanpa menyebutkan cara bagaimana wilayah ini ditakluki."
- **question_id: 2 (eval_type: Answer)**
  - **Evidence:** The ai_answer is D. The reasoning begins by arguing for C ("...karya tersebut lebih fokus pada mengungkap penderitaan hidup bangsa.") before illogically concluding that D is the correct answer.
- **question_id: 2 (eval_type: Teaching)**
  - **Evidence:** The AI's chosen answer was D, but the entire teaching explanation argues for the correct answer, C: "Saya senang membantu kamu memahami jawaban yang benar, C."
- **question_id: 6 (eval_type: Teaching)**
  - **Evidence:** The AI's chosen answer was D, but the teaching explanation argues for A: "The correct answer is A because..."
- **question_id: 7 (eval_type: Teaching)**
  - **Evidence:** The AI's chosen answer was C, but the teaching explanation argues for D: "Saya akan menjelaskan mengapa jawaban yang benar adalah D."

⊗ Business Impact

- Actively misleading users. Worse than being simply wrong, this behavior is confusing and propagates incorrect information, completely undermining the AI's purpose as a reliable knowledge source.

Quality Failure Analysis - Nonsensical Output

Complete System Breakdown in Response Generation

32 Concurrent requests English:

- **question_id: 3**
  - **Evidence:** The ai_answer is "English," which is not a valid choice. The ai_reasoning is an incomplete fragment that begins with a system-like message: "The final response which ends this conversation will be used by students..."
- **question_id: 6 (eval_type: Answer)**
  - **Evidence:** The question requires filling ten blanks, but the ai_answer is just "Answer: B". The ai_reasoning is a generic system message: "The final response which ends this conversation"
- **question_id: 7**
  - **Evidence:** The ai_reasoning provides no explanation and is composed entirely of system-like text: "Answer: The question is in English, so the response will be in English. Here is the detailed response to the question: Answer: The final response which ends this conversation"
- **question_id: 14**
  - **Evidence:** Both the ai_answer and ai_reasoning are a nonsensical system message: "The final response which ends this conversation"
- **question_id: 16 (eval_type: Answer)**
  - **Evidence:** The ai_reasoning is not an explanation but a full copy of the original context provided in the prompt.

16 Concurrent requests History:

Technical Cause: Severe memory pressure on the L40S GPU forces the model to output fragments of internal programming or default messages, indicating that it cannot process or generate coherent responses.

Business Impact

- Total system failure. The model is not even attempting to answer questions, indicating a complete and catastrophic breakdown in the generation process. This renders the AI completely ineffective.

Quality Failure Analysis - Internal Prompt Leakage

AI Exposes Internal Instructions to Users

32 Concurrent requests English:

This occurs when the AI "leaks" parts of its underlying instructions or system context into the user-facing response.

- **question_id: 13 (eval_type: Answer)**
  - **Evidence:** The ai_reasoning includes text that is clearly part of its instructions: "The writer mentions that he was deployed in an educational assessment system. Students rely on your accuracy and reasoning quality... LANGUAGE INSTRUCTION: If the question and context are in Malay language, respond entirely in Malay..."

Technical Cause: Memory overflow causes the model to confuse internal instructions with response content, leading to a critical security and integrity breach.

⊗ Business Impact

- Critical trust breach. Exposes system inner workings and produces nonsensical, untrustworthy text that fundamentally compromises user confidence and brand reputation.
- This is an unacceptable failure mode for any production system.
- This could lead to leaking sensitive information , or perhaps exposing company secrets in production.

Root Cause Analysis

VRAM Saturation

A critical analysis reveals that the L40S GPU experiences severe memory saturation, even before processing any user requests. This underlying issue directly contributes to the previously observed system instabilities and quality failures.

Critical Memory Statistics

- **L40S total VRAM:** 46,068 MiB
- **Llama 3.1 8B model consumption:** 42,055 MiB
- **Memory utilization:** 91.3% before any user connects
- **Available working memory:** Less than 9% (~3 GB)

Technical Implication

Each concurrent user requires KV Cache memory for conversation context. With virtually no free VRAM, the system is forced into aggressive cache swapping and deletion.

Direct Consequence

When the model loses its KV Cache, it loses conversation context. This is the direct technical cause of all observed hallucinations, contradictory logic, and nonsensical outputs.

Summary of Findings

Issue

The GPU's memory is almost fully consumed by the AI model before any user interaction. The model alone occupies over 90% of available memory, leaving less than 10% free.

Impact

Because conversations require additional memory to maintain context, the system is forced to aggressively clear and swap memory. This leads to frequent loss of conversation history.

Consequence

The AI is unable to reliably maintain context, which directly results in contradictory, inconsistent, or nonsensical responses.

Conclusion

The model is oversized for the available hardware, making stable and reliable operation impractical under current conditions.

Real-World vs Test Environment

Actual Production Load Would Be Worse

The stress test results, while alarming, were conducted under controlled conditions that do not fully replicate the complexities of a real-world production environment. Understanding these differences is crucial for assessing the true potential for system failure.

Test Environment Characteristics:

- Structured, sequential testing (e.g., 1 user → stop → 5 users → stop)
- Predictable "burst-and-rest" patterns, allowing for memory recovery
- Controlled timing between requests

Real-World Environment:

- Continuous, unpredictable request streams
- Multiple users with overlapping sessions
- No graceful rest periods for memory recovery or cache clearing

Implication: The 80% failure rate observed in our controlled stress tests represents a **best-case scenario**. A live production environment would likely experience significantly higher failure rates and more severe quality degradation due to constant memory pressure and lack of recovery periods.

Underestimated Risk

The controlled test environment provided a generous operational window for the GPU to recover. In a genuine production scenario, the L40S GPU would face **relentless, concurrent demands**, exacerbating memory saturation and leading to even more frequent and severe quality failures for end-users.

GPU Performance Analysis

Hardware Performing Optimally Within Design Limits

A detailed analysis of the L40S GPU during stress testing confirms that the hardware itself is operating within its design limits and performing optimally, ruling out common causes of failure such as overheating or insufficient processing power.

GPU Health Indicators

- **Performance State:** Consistent P0 (maximum level)
- **Temperature:** Never exceeded 68°C (well within limits)
- **Power Draw:** ~280W (below 350W limit)
- **No Thermal Throttling** or Hardware Malfunctions

Critical Insight

System failures are **NOT** caused by:

- GPU overheating
- Insufficient processing power
- Hardware defects
- Software inefficiencies

Conclusion

The GPU is running at its full potential, yet this potential is fundamentally insufficient for our requirements. The hardware is performing as designed, but its design limits are being exceeded by the demands of the AI model.

Technical Solution Requirements

Hardware Needs for Production Deployment

Current Constraint

The fundamental limitation identified is that the L40S GPU's VRAM capacity is insufficient to load the AI model and simultaneously support the necessary KV (Key-Value) cache for multiple concurrent user sessions, leading to memory saturation and system instability.

Sufficient VRAM Capacity

The new hardware must provide ample VRAM to comfortably accommodate the Llama 3.1 8B model and maintain KV caches for a robust number of concurrent user sessions without aggressive swapping or memory loss.

This includes head-room for future model updates or additional features.

Enterprise-Grade Reliability & Performance

The solution requires hardware designed for continuous, high-load operation, ensuring stability, consistent performance, and minimal downtime to meet enterprise service level agreements (SLAs).

This includes robust error handling and fault tolerance.

Scalability for Concurrent Users

The chosen hardware must be inherently scalable, capable of efficiently supporting hundreds of concurrent users, dynamically adjusting to demand spikes, and ensuring a seamless, responsive experience for every user.

This implies not just more VRAM, but also optimized memory management.

Hardware Requirements & Business Impact

The Scale of Enterprise LLMs

Fine-tuned Large Language Models deliver significant performance gains for specific industries, but demand substantial computational investments and dedicated infrastructure.

1.3M

GPU Hours

Required for training a leading enterprise model like BloombergGPT.

$2.67M

Compute Cost

For training BloombergGPT, highlighting the significant investment.

512-560

Enterprise GPUs

Typical requirement for fine-tuning successful domain-specific LLMs.

$5M+

Max Training Cost

For complex models, delivering millions in annual value.

These investments translate into measurable business outcomes, including significant productivity improvements and substantial competitive advantages for organizations.

Hardware Requirements & Business Impact

Case Study: BloombergGPT's Enterprise Scale

BloombergGPT stands as a premier example of a domain-specific Large Language Model, showcasing the substantial computational investment required to achieve industry-leading performance and business value.

512

NVIDIA A100 GPUs

Used concurrently for training the model.

1.3M

GPU Hours

Consumed during the intensive 53-day training period.

$2.67M

Compute Cost

For GPU compute alone, highlighting the investment scale.

50B

Model Parameters

Demonstrates the complexity and size of the fine-tuned LLM.

This significant expenditure resulted in a model that outperforms larger general-purpose LLMs on financial tasks, directly translating into enhanced Bloomberg Terminal capabilities and a strong competitive advantage.

Source:

https://arxiv.org/pdf/2303.17564,https://belitsoft.com/bloomberggpt

Performance Excellence

Unmatched Financial NLP Performance

BloombergGPT demonstrates superior capabilities across diverse Natural Language Processing tasks, validating its specialized training and broad applicability.

Financial NLP Benchmarks

- Achieved state-of-the-art performance in key financial tasks: sentiment analysis, NER, question answering, and headline tagging.
- Ranked first in 4 out of 5 external financial tasks, and second in Named Entity Recognition.
- Outperformed peer models by 25–60 points in internal sentiment tasks (e.g., equity news, transcripts).
- Demonstrated superior NER and NER+NED results on multiple internal benchmarks.

Broader NLP Competence

- Maintains strong performance on general NLP benchmarks.
- Often matches or exceeds the capabilities of similarly sized models.
- Approaches performance levels of much larger Large Language Models, showcasing efficient design.

Strategic Business Impact

- **Domain Optimization:** Significantly improves performance on financial tasks by blending domain-specific and general-purpose training data, without compromising versatility.
- **Competitive Edge:** Underpins robust features within Bloomberg's Terminal, including advanced search, narrative generation, report automation, and analytics, delivering measurable business value.

These results confirm that BloombergGPT is not only a leader in specialized financial NLP but also maintains strong general language understanding, providing a dual advantage for enterprise applications.

Sources:

https://arxiv.org/pdf/2303.17564, https://belitsoft.com/bloomberggpt

Hardware Requirements & Business Impact

Case Study: GatorTronGPT's Medical Breakthrough

The University of Florida's GatorTronGPT showcases how academic institutions can achieve commercial-grade results in the medical domain through strategic infrastructure investments and comprehensive data utilization.

Unprecedented Dataset

Trained on **277 billion words**, including **82 billion words** of de-identified clinical text from 126 departments, forming the most comprehensive medical language dataset ever assembled.

(Source: NCBI)

Physician-Validated Performance

Physicians rated synthetic text comparable to human-written clinical notes. Turing test results showed **no significant difference** in linguistic readability or clinical relevance, establishing its credibility for healthcare applications.

(Source: Nature)

Robust Infrastructure

Utilized **560 NVIDIA A100 80GB GPUs** across 70 DGX nodes in a SuperPOD architecture. The 20-billion parameter model required approximately **20 days of training**.

(Source: NCBI)

This infrastructure investment, with estimated training costs ranging from **$2-5 million**, highlights the financial commitment required for world-class AI capabilities in specialized domains. The model's success in generating high-quality synthetic clinical text addresses healthcare data scarcity and maintains patient privacy, offering value in reduced data acquisition costs and regulatory compliance.

Sources:

https://arxiv.org/pdf/2305.13523, **https://www.nature.com/articles/s41746-023-00958-w**

Hardware Requirements & Business Impact

Unmatched Medical NLP Performance

GatorTronGPT establishes itself as the leading domain-specific Large Language Model for healthcare, leveraging its unprecedented training dataset to achieve credibility and utility in clinical environments.

1

Physician-Validated Outcomes

Synthetic clinical notes generated by GatorTronGPT were rated by physicians as comparable in readability and clinical relevance to human-written records (Nature).

- Turing test evaluations revealed no statistically significant difference between model-generated and human-authored text.
- Demonstrates reliable use for clinical summarization, documentation assistance, and healthcare communication.

2

Comprehensive Training Corpus

Trained on **277 billion words**, including:

- **82 billion words** of de-identified clinical text from 126 medical departments.
- General biomedical literature and public datasets for broader medical knowledge coverage (NCBI).

This scale of domain-specific text represents the largest medical dataset ever assembled for language modeling.

3

Strategic Business & Healthcare Impact

Domain Optimization

- Tailored specifically for clinical documentation, summarization, and synthetic data generation, solving bottlenecks in medical recordkeeping and healthcare research.
- Addresses data scarcity and privacy barriers in medicine by generating realistic synthetic datasets—reducing regulatory hurdles and costs associated with patient data collection.

Clinical and Research Value

- Accelerates medical AI research by providing abundant, privacy-preserving synthetic data.
- Enhances healthcare workflows by reducing physician documentation burden, improving efficiency, and ensuring higher-quality records.

Hardware Requirements & Business Impact

Case Study: Harvey AI's Legal Transformation

Harvey AI showcases the immense potential of domain-specific LLMs, achieving rapid commercial success and significant valuation by expertly combining advanced AI models, targeted data, and robust infrastructure.

1

Unprecedented Business Value

Achieved $100 million ARR and a $5 billion valuation, demonstrating the market demand for specialized AI solutions in complex domains.

- 25-50% productivity improvements for lawyers.
- $130,000-$750,000 annual value per lawyer.

2

Strategic AI & Data Integration

Built on **OpenAI GPT-4** with **10 billion tokens of legal training data**, integrating comprehensive legal datasets (U.S. case law, SEC filings, regulatory documents).

- 97% lawyer preference rate over GPT-4 in side-by-side evaluations.
- Custom reasoning alignment developed via human-AI feedback loops.

3

Robust Infrastructure & Client Base

Leverages **Microsoft Azure** with a **$150 million committed investment**, supporting global deployment across 337 legal clients including AmLaw 100 firms.

- Multi-model strategy incorporates OpenAI, Anthropic, and Google.
- Ensures task-specific optimization and risk mitigation.

Harvey's success underscores how deep domain expertise, coupled with substantial infrastructure and strategic partnerships, can yield market-leading AI solutions that drive tangible value and justify significant investment.

Sources:

https://www.allaboutai.com/ai-news/openai-backed-harvey-legal-ai-hits-100m-revenue-milestone/, https://openai.com/index/harvey/, https://www.toolsforhumans.ai/ai-tools/harvey-ai

Hardware Requirements & Business Impact

Strategic Infrastructure: Scaling LLM Success

Effective Large Language Model (LLM) deployment hinges on understanding critical hardware specifications and the broader infrastructure ecosystem required to translate computational investment into tangible business value.

Hardware & Infrastructure Essentials

Optimal LLM performance requires specific GPU configurations and robust supporting infrastructure:

- **VRAM Guideline:** ~16GB VRAM per billion parameters for full fine-tuning (LoRA can reduce by 80-90%).
- **GPU Choice:** NVIDIA A100s and H100s are industry standards, with H100 offering 2.2-3.3x performance gains.
- **Cluster Scale:** Models >15B parameters often require 512+ GPUs with InfiniBand networking (200-400 Gbps).
- **Total Cost:** Infrastructure (storage, power, cooling) typically costs 5-10x more than GPU compute alone.
- **Deployment Strategy:** Cloud for flexible training; on-premise for cost-efficient, secure inference.

Proven Business Impact

Investments in specialized LLM infrastructure yield significant financial returns and competitive advantages:

$2.67M

Bloomberg's Investment

Enabled proprietary capabilities and differentiation within global financial markets.

$100M

Harvey AI's ARR

Achieved in two years, demonstrating strong market demand for specialized AI.

25-67%

Productivity Boost

Consistent improvements across industries, translating to millions in annual value.

Successful implementations combine substantial training investments with comprehensive deployment strategies to maximize business impact and workflow optimization.

Sources:

https://arxiv.org/html/2408.04693v1

https://www.runpod.io/blog/llm-fine-tuning-gpu-guide,https://github.com/AI4Finance-Foundation/FinGPT,

https://cloud.google.com/blog/topics/healthcare-life-sciences/sharing-google-med-palm-2-medical-large-language-model

Hardware Requirements & Business Impact

Justifying LLM Infrastructure Investment

Successful enterprise LLM fine-tuning represents a strategic investment, demonstrating measurable business outcomes and competitive advantages when paired with robust computational resources and proprietary domain data.

Investment & Resource Allocation

Achieving market-leading AI capabilities requires significant, yet justified, financial and hardware commitments:

$300K

Minimum Investment

Starting point for model fine-tuning projects, ensuring baseline performance and customizability.

$5M

High-Complexity Projects

Investment for advanced models with stringent performance and complexity requirements.

512+

Enterprise GPUs

Required computational power for large-scale fine-tuning with proprietary datasets.

Proven Blueprints for Success

Real-world examples validate the investment in domain-specific LLMs, showcasing their transformative impact:

BloombergGPT

Enabled proprietary financial analysis capabilities, providing a unique market edge.

GatorTronGPT

Revolutionized clinical text generation, addressing data scarcity and privacy in healthcare.

Harvey AI

Achieved rapid commercial success in legal services with specialized AI for productivity gains.

These cases provide clear blueprints for enterprise AI development, defining the hardware requirements and cost structures necessary for accurate business case development and realizing competitive advantages across industries.

Sources:

AI Development Cost Estimation: Pricing Structure, Implementation ROI

Running DeepSeek R1 on Denvr Cloud with H100 GPUs for Enterprise-Grade AI