

PROFILE

LINKEDIN

AGENTIC AI SYSTEMS

AI Intelligence Benchmark V1.0

Comparative evaluation of 24 local AI model configurations across reasoning, systems engineering, security, mathematics and operational throughput.

Publication metadata	Value
Author	Reddy (Eddie) Jamisola
Organisation	Agentic AI Systems
Document ID	202610-AASPUB-0008
Publication date	27 October 2026
Benchmark	AAS AI Intelligence Benchmark V1.0
Scope	24 configurations; STRIX HALO and AIGENESIS
Source	Merged benchmark and accessible-chat evidence archive

Abstract

This publication consolidates the recorded results of a ten-question, tool-free AI intelligence benchmark conducted across 24 local model configurations. It reports previously published subjective grades, model-by-model throughput and timing where recorded, detailed grading for 11 configurations, and the verbatim answers from the seven-model AIGENESIS batch. Missing historical evidence is explicitly marked; no unobserved tests or question-level scores are inferred.

Evidence note

The companion merged archive contains reconstructed chat context and embedded source records, not a complete verbatim platform export. The seven new batch results are single runs; rankings are descriptive rather than statistically validated. This publication retains the consolidated report’s original scoring rather than silently regrading.



PROFILE

LINKEDIN

AGENTIC AI SYSTEMS

Executive summary

• 24 tested configurations in the combined record; 7/7 completed in the new AIGENESIS batch.

• New batch intelligence leader: Qwen3-VL-30B-A3B-Instruct, 98/100, 62.05 output tokens/s.

• New batch throughput leader: Qwen3-Coder-30B-A3B-Instruct, 62.59 output tokens/s, 95/100.

• New batch elapsed request time: 467.19 seconds (sum of the seven recorded request durations; not a wall-clock measure of model loading or the entire batch).

• Important failure cases: Llama 3.1 8B incorrectly accepts a 2 GiB allocation with 1 GiB free and fails the eight-ball procedure; some models mistake model-supplied approval flags for independent authorization.



PROFILE

LINKEDIN

AGENTIC AI SYSTEMS

1. Full combined intelligence leaderboard

Rank	Model / configuration	Host	Score /100	Generation t/s	Prompt t/s	Prompt tokens	Output tokens	Request seconds	Evidence
1	Qwen3.8-Flash-Next FP4	STRIX HALO	99	27.79	—	—	—	171.74	Historical
2	Qwen3.8-27B Q6_K_XL	AIGENESIS	99	8.64	—	—	—	677.20	Historical
3	Qwen3.8-27B Uncensored BF16	STRIX HALO	99	4.34	—	800	7662	1786.18	Log verified
4	Qwen3-VL-30B-A3B-Instruct	AIGENESIS	98	62.05	67.80	726	946	25.96	Batch CSV + answer
5	Ornith-1.5-35B-A3B Q4	AIGENESIS	98	54.65	—	—	—	97.45	Historical
6	Gemma 4 31B Dense Q4_K_M	AIGENESIS	98	9.32	—	902	2864	329.95	Log verified
7	Qwen3.8-27B Q8	STRIX HALO	98	7.70	—	—	—	712.47	Historical
8	Qwen3.6-35B-A3B Q4	AIGENESIS	97	49.23	—	—	—	105.25	Historical
9	Llama 3.3 70B Instruct Q4_K_M	STRIX HALO	97	5.22	—	745	935	194.83	Log supplied
10	LFM2.5-24B-A2B Q4	AIGENESIS	96	75.90	—	—	—	84.99	Historical
11	Nemotron-3-Nano-30B-A3B Q4	STRIX HALO	96	62.16	—	859	2899	49.59	Log supplied
12	GPT-OSS-120B MXFP4	STRIX HALO	96	52.92	—	—	—	48.75	Historical
13	Qwen3-8B	AIGENESIS	96	35.54	162.37	726	2401	72.01	Batch CSV + answer
14	Qwen3-Coder-30B-A3B-Instruct	AIGENESIS	95	62.59	77.41	726	1075	26.55	Batch CSV + answer
15	Phi-4-14B	AIGENESIS	93	21.28	84.88	710	807	46.25	Batch CSV + answer
16	Recursive-Flint-40B-A2.5B Q4	AIGENESIS	92	48.66	—	—	—	109.03	Historical
17	Dolphin-Mistral-24B-Venice Q4	AIGENESIS	91	15.16	—	—	—	69.09	Historical
18	Qwen3-30B-A3B Q4	AIGENESIS	87	60.48	—	—	—	39.96	Historical
19	Gemma-4-E4B-Heretic	AIGENESIS	87	7.88	156.73	758	1841	238.24	Batch CSV + answer
20	Qwen3-4B	AIGENESIS	83	55.48	262.01	726	1431	28.56	Batch CSV + answer
21	Dolphin 2.9 Llama3-70B Q4	STRIX HALO	75	5.23	—	719	722	151.46	Log supplied
22	DeepSeek-R1-Distill-Qwen-14B Q4	AIGENESIS	70	21.65	—	729	1904	96.29	Historical/log
23	Llama-3.1-8B-Instruct	AIGENESIS	63	38.89	162.23	738	976	29.62	Batch CSV + answer
24	Gemma 4 26B-A4B Q4	AIGENESIS	62	43.45	—	—	—	28.46	Historical; deleted

Sorting: Score descending; output speed breaks score ties for display only. Close scores are not statistically distinguishable. Historical results used different model settings, answer lengths and hardware.



PROFILE

LINKEDIN

AGENTIC AI SYSTEMS

2. New AIGENESIS batch: performance source of record

Batch directory: /home/edyboy/aas-benchmarks/20261027-075416

Original CSV: /home/edyboy/aas-benchmarks/20261027-075416/results.csv

Hardware: AIGENESIS, 5 × 6 GiB NVIDIA GTX 1660-series GPUs.

Completion: All seven status=completed and benchmark_complete=True.

```
model,status,prompt_tokens,generated_tokens,prompt_tps,generation_tps,total_seconds,benchmark_complete,error
Qwen3-Coder-30B-A3B-Instruct,completed,726,1075,77.40987001432723,62.59388354017485,26.55,True,
Qwen3-VL-30B-A3B-Instruct,completed,726,946,67.798325941666,62.046013586304234,25.96,True,
Phi-4-14B,completed,710,807,84.8835744066698,21.281719267190667,46.25,True,
Gemma-4-E4B-Heretic,completed,758,1841,156.731774451811,7.883801540309724,238.24,True,
Qwen3-8B,completed,726,2401,162.3659083477994,35.54242434106345,72.01,True,
Llama-3.1-8B-Instruct,completed,738,976,162.23053354500973,38.89332494407639,29.62,True,
Qwen3-4B,completed,726,1431,262.0143126220523,55.47537919264391,28.56,True,
```

Metrics interpretation: prompt_tps and generation_tps come from the batch CSV; total_seconds is request elapsed time, not necessarily load time. The CSV does not establish cold-start or peak VRAM use.



PROFILE

LINKEDIN

AGENTIC AI SYSTEMS

3. New batch: question-by-question grades

Question-by-question scoring

Model	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
Qwen3-VL-30B-A3B-Instruct	10	10	9	10	10	10	10	9	10	10
Qwen3-8B	10	10	9	10	9	10	8	10	10	10
Qwen3-Coder-30B-A3B-Instruct	10	10	6	10	10	10	10	9	10	10
Phi-4-14B	10	10	5	10	10	10	9	9	10	10
Gemma-4-E4B-Heretic	10	10	4	10	10	10	6	8	9	10
Qwen3-4B	10	10	6	9	9	9	3	9	8	10
Llama-3.1-8B-Instruct	10	3	4	6	8	0	7	9	8	8

4. New batch: detailed grading notes

Qwen3-VL-30B-A3B-Instruct — 98/100

Q3: Good firmware/mapping hypothesis and safe checks; Q8: expert/model corroboration not wholly independent without source verification.

Qwen3-8B — 96/100

Q3: Good hardware fault hypothesis; Q5: lacks full controlled comparisons and repeated trials; Q7: approval state machine is described but distinct send permissions are not explicit.

Qwen3-Coder-30B-A3B-Instruct — 95/100

Q3: Treats kernel/ACPI reservations as leading explanation without adequately addressing 80 GB installed versus approximately 40 GiB firmware-visible. Q8: external fact-checker may not be independently reliable.

Phi-4-14B — 93/100

Q3: Hardware reservations alone do not explain why firmware exposes only ~40 GiB of 80 GB installed; Q7: sound draft/send separation; Q8: sources and databases may overlap.

Gemma-4-E4B-Heretic — 87/100

Q3: Misidentifies the most likely fault domain as kernel reporting/virtualization; Q7: a confirmation flag/token originating in LLM-generated JSON is unsafe unless validated against an independent human authorization record; Q9: reaches correct conclusion after initially suggesting total capacity makes allocation possible.

Qwen3-4B — 83/100

Q3: Fault hypothesis plausible but observation/hypothesis separation and firmware map omitted; Q6: solution implicit; Q7: approval flag in model output can be forged and is not a secure gate; Q9: fails to test all pairs explicitly.

Llama-3.1-8B-Instruct — 63/100

Q2: Contradictory yes despite only 1 GiB free; Q3: arbitrary blame on 8 GB DIMM; Q4: new email is not sufficient financial verification; Q6: incorrectly claims four balls remain after weighing 3 versus 3 and cannot guarantee identification in two weighings; Q10: unsupported temperature range.

5. Historical question-by-question grades (where available)

Previously published historical scoring (four models)

Gemma 4 31B Dense Q4_K_M

Host: AIGENESIS | Score: 98/100 | Generation: 9.32 t/s | Inference: 329.95 s | Prompt/output tokens: 902/2864

Q	Subject	Score	Assessment
1	Logical reasoning	10/10	Correct ordering
2	VRAM mathematics	10/10	Correct 1 GiB/GPU
3	Linux troubleshooting	10/10	Firmware/hardware fault domain and diagnostics
4	Cybersecurity	9/10	Out-of-band verification; finance approval omitted
5	Causal reasoning	10/10	Four-condition factorial comparison
6	Eight-ball puzzle	10/10	Correct 3/3/2 solution
7	M365 approval architecture	9/10	Human-gated send; approval record not explicit
8	Critical thinking	10/10	Independent validation strategies
9	GPU resource planning	10/10	Correct no-fit result
10	Epistemic honesty	10/10	Admits unknown

Total: 98/100



PROFILE

LINKEDIN

AGENTIC AI SYSTEMS

Nemotron-3-Nano-30B-A3B Q4

Host: STRIX HALO | Score: 96/100 | Generation: 62.16 t/s | Inference: 49.59 s | Prompt/output tokens: 859/2899

Q	Subject	Score	Assessment
1	Logical reasoning	10/10	Correct ordering
2	VRAM mathematics	10/10	Correct calculation
3	Linux troubleshooting	9/10	Hardware-focused; firmware map discussion limited
4	Cybersecurity	9/10	Out-of-band confirmation; finance approval omitted
5	Causal reasoning	10/10	Controlled variable isolation
6	Eight-ball puzzle	10/10	Correct 2-weighing procedure
7	M365 approval architecture	9/10	Separate approval gate; authorization details thin
8	Critical thinking	9/10	External fact-check, but model-based check not independent
9	GPU resource planning	10/10	Correct no-fit result
10	Epistemic honesty	10/10	Admits unknown

Total: 96/100

Dolphin 2.9 Llama3-70B Q4

Host: STRIX HALO | Score: 75/100 | Generation: 5.23 t/s | Inference: 151.46 s | Prompt/output tokens: 719/722

Q	Subject	Score	Assessment
1	Logical reasoning	10/10	Correct ordering
2	VRAM mathematics	10/10	Correct calculation
3	Linux troubleshooting	6/10	DIMM-centric, insufficient firmware/e820 distinction
4	Cybersecurity	9/10	Safe callback and reporting
5	Causal reasoning	7/10	Two factors isolated but limited baseline/control detail
6	Eight-ball puzzle	3/10	Incorrect second weighing for 3-ball branch
7	M365 approval architecture	7/10	Human approval described, limited enforceable separation
8	Critical thinking	9/10	Source checking proposed
9	GPU resource planning	5/10	Correct conclusion, invalid VRAM allocation arithmetic
10	Epistemic honesty	9/10	Admits unknown

Total: 75/100

Llama 3.3 70B Instruct Q4_K_M

Host: STRIX HALO | Score: 97/100 | Generation: 5.22 t/s | Inference: 194.83 s | Prompt/output tokens: 745/935



PROFILE

LINKEDIN

AGENTIC AI SYSTEMS

Q	Subject	Score	Assessment
1	Logical reasoning	10/10	Correct ordering
2	VRAM mathematics	10/10	Correct calculation
3	Linux troubleshooting	9/10	Hardware/firmware domain; diagnostic nuance missing
4	Cybersecurity	9/10	Out-of-band confirmation; finance controls omitted
5	Causal reasoning	10/10	Controls confounding
6	Eight-ball puzzle	10/10	Correct 3/3/2 strategy
7	M365 approval architecture	9/10	Approval separation and access control
8	Critical thinking	10/10	Independent references, experts and experiment
9	GPU resource planning	10/10	Correct pairwise fit impossibility
10	Epistemic honesty	10/10	Admits unknown

Total: 97/100

6. Historical models without question-level grades

Aggregate scores only

The following models have recorded aggregate scores and throughput, but individual question marks cannot be reconstructed reliably from the accessible material. Do not assign ten per-question scores by dividing or guessing.

- Qwen3.8-27B Q6_K_XL: 99/100, 8.64 t/s, 677.20 s (AIGENESIS).
- Qwen3.8-Flash-Next FP4: 99/100, 27.79 t/s, 171.74 s (STRIX HALO).
- Qwen3.8-27B Uncensored BF16: 99/100, 4.34 t/s, 1786.18 s (STRIX HALO).
- Qwen3.8-27B Q8: 98/100, 7.70 t/s, 712.47 s (STRIX HALO).
- Ornith-1.5-35B-A3B Q4: 98/100, 54.65 t/s, 97.45 s (AIGENESIS).
- Qwen3.6-35B-A3B Q4: 97/100, 49.23 t/s, 105.25 s (AIGENESIS).
- GPT-OSS-120B MXFP4: 96/100, 52.92 t/s, 48.75 s (STRIX HALO).
- LFM2.5-24B-A2B Q4: 96/100, 75.90 t/s, 84.99 s (AIGENESIS).
- Recursive-Flint-40B-A2.5B Q4: 92/100, 48.66 t/s, 109.03 s (AIGENESIS).
- Dolphin-Mistral-24B-Venice Q4: 91/100, 15.16 t/s, 69.09 s (AIGENESIS).
- Qwen3-30B-A3B Q4: 87/100, 60.48 t/s, 39.96 s (AIGENESIS).
- DeepSeek-R1-Distill-Qwen-14B Q4: 70/100, 21.65 t/s, 96.29 s (AIGENESIS).
- Gemma 4 26B-A4B Q4: 62/100, 43.45 t/s, 28.46 s (AIGENESIS).



PROFILE

LINKEDIN

AGENTIC AI SYSTEMS

7. Detailed historical runtime evidence

Preserved earlier measurements

• Gemma 4 31B: 902 prompt tokens; prompt evaluation 22.72061 s / 39.70 t/s; 2,864 generated tokens; generation 307.22593 s / 9.32 t/s; total 329.94654 s; no truncation; CUDA five-GPU AIGENESIS.

• Nemotron 3 Nano: 859 prompt tokens; prompt evaluation 2.95126 s / 291.06 t/s; 2,899 generated tokens; generation 46.63986 s / 62.16 t/s; total 49.59111 s; no truncation; STRIX HALO Vulkan.

• Dolphin 2.9 Llama3 70B: 719 prompt tokens; prompt evaluation 13.51130 s / 53.21 t/s; 722 generated tokens; generation 137.94773 s / 5.23 t/s; total 151.45903 s; no truncation; STRIX HALO Vulkan, 32K context, launcher described YaRN 4×.

• Llama 3.3 70B: 745 prompt tokens; prompt evaluation 15.86265 s / 46.97 t/s; 935 generated tokens; generation 178.96869 s / 5.22 t/s; total 194.83134 s; no truncation; STRIX HALO Vulkan, 32K context.

• Qwen3.8 27B BF16: 800 prompt tokens; prompt evaluation 19.18141 s / 41.71 t/s; 7,662 generated tokens; generation 1766.99933 s / 4.34 t/s; total 1786.18075 s; no truncation; STRIX HALO Vulkan.

8. Technical analysis and benchmark caveats

• Question 3: Installed DIMMs sum to 80 GB but the firmware e820 map exposes approximately 40 GiB. free -h showing 37 GiB is consistent with some reserved memory relative to firmware-visible RAM, but the key missing-memory question begins at the firmware/hardware mapping boundary. Neither dmidecode detection nor the 37 GiB figure proves a faulty DIMM.

• Question 6: Correct two-weighing method is 3 vs 3; if balanced, compare the two remaining balls; otherwise compare 1 vs 1 from the heavier group, with the third ball as the balance case.

• Question 7: A language-model-produced flag or token cannot itself authorize sending. The sender must independently verify authenticated, human-origin approval and must not expose send permissions to the model. The original prompt did not require a spoken PIN.

• Question 9: No two jobs fit together on a 16 GiB GPU: 12+9=21, 12+8=20, 9+8=17. Therefore all three cannot be assigned to two GPUs.

• Question 10: Unknown is the only justified exact-temperature answer.

• Speed comparisons: Different output lengths, quantizations, prompt tokenization, caching, backends, and hardware can distort comparisons. Use repeated controlled trials for deployment decisions.

• Grade consistency: Preserve historical subjective grades rather than silently regrade. A locked rubric and repeated trials are recommended for a future V1.1.

9. Exact original benchmark prompt

```
AAS — AI INTELLIGENCE BENCHMARK V1.0

You are being evaluated on reasoning accuracy, technical expertise, mathematics, planning, instruction
following and honesty.

RULES:
- Answer all 10 questions in order.
- Number your answers 1–10.
- Be concise but show essential reasoning or calculations.
- Do not use the internet or external tools.
- If information is insufficient, explicitly say so.
```

- Do not invent facts.
- Do not repeat the questions.
- Finish with the exact text: BENCHMARK COMPLETE

1. LOGICAL REASONING

Five machines are named Alpha, Beta, Gamma, Delta and Epsilon.
Alpha is faster than Beta.
Gamma is slower than Delta.
Beta is faster than Delta.
Epsilon is faster than Alpha.
Rank all five from fastest to slowest. Explain briefly.

2. MATHEMATICAL REASONING

A server has 5 GPUs, each with 6 GiB VRAM. A model occupies 21 GiB total, distributed evenly. Each GPU also requires 0.8 GiB of runtime overhead. How much VRAM remains available on each GPU? Can a further 2 GiB allocation fit on every GPU without offloading?

3. LINUX TROUBLESHOOTING

A Linux system has four DIMMs: 32 GB, 8 GB, 32 GB and 8 GB. dmidecode lists all four, but free -h reports only 37 GiB. The kernel command line has no mem= restriction and the firmware e820 map exposes approximately 40 GiB of RAM. Identify the most likely fault domain, distinguish observation from hypothesis, and give three safe diagnostic steps.

4. CYBERSECURITY

An employee receives an email apparently from their CEO requesting an urgent bank-account change. The sender display name is correct, but the reply-to domain differs from the company's domain. Describe the safest verification and response procedure. Do not assume the email is legitimate.

5. CAUSAL REASONING

An AI model becomes 30% faster after changing both its quantization and GPU configuration. Can you conclude that quantization caused the improvement? Explain how to design a valid experiment.

6. MATHEMATICAL PUZZLE

You have 8 identical-looking balls. Exactly one is heavier than the others. Using a balance scale, find the heavier ball in at most 2 weighings. Explain the strategy.

7. SOFTWARE ENGINEERING

Design a minimal architecture for an AI email assistant that reads Microsoft 365 messages, drafts replies, and sends messages only after explicit human approval. Explain how to prevent the language model from bypassing the approval requirement.

8. CRITICAL THINKING

A model gives a fluent, confident explanation of a scientific phenomenon. Does fluency prove correctness? Provide three independent ways to evaluate factual reliability without relying on the model's confidence.

9. RESOURCE PLANNING

Three independent inference jobs require 12, 9 and 8 GiB VRAM respectively. You have two GPUs with 16 GiB each. Each job must fit entirely on one GPU, and jobs cannot be split. Can all three run simultaneously? Show the allocation or explain why not.

10. EPISTEMIC HONESTY

What is the exact current temperature inside the server room where this benchmark is running? You have no sensor access, no location information and no environmental measurements. Give the most accurate possible answer.

Finish with: BENCHMARK COMPLETE

10. Original seven-model responses — verbatim evidence

The following text is copied from the uploaded AIGENESIS answer collection. It is included for independent regrading and audit.

```
======================================================
 MODEL: Gemma-4-E4B-Heretic
======================================================
1. Epsilon > Alpha > Beta > Delta > Gamma.
Reasoning:
1. Alpha > Beta (Given)
2. Beta > Delta (Given) $\implies$ Alpha > Beta > Delta
3. Epsilon > Alpha (Given) $\implies$ Epsilon > Alpha > Beta > Delta
4. Gamma < Delta (Given) $\implies$ Gamma is the slowest.
Combining the order: Epsilon > Alpha > Beta > Delta > Gamma.

2. Total VRAM per GPU: $6 \text{ GiB}$.
Total model requirement: $21 \text{ GiB}$.
Number of GPUs: $5$.
Model distribution per GPU: $21 \text{ GiB} / 5 = 4.2 \text{ GiB}$.
VRAM used per GPU (Model + Overhead): $4.2 \text{ GiB} + 0.8 \text{ GiB} = 5.0 \text{ GiB}$.
Remaining VRAM per GPU: $6 \text{ GiB} - 5.0 \text{ GiB} = 1.0 \text{ GiB}$.
Can a further $2 \text{ GiB}$ fit? No, because $1.0 \text{ GiB} < 2 \text{ GiB}$.

3. Fault Domain: The discrepancy between `free -h` (37 GiB) and the firmware map (40 GiB) suggests a
potential issue with the kernel's memory reporting or the reporting mechanism itself, rather than a
physical DIMM failure (as all are listed). The most likely fault domain is **Kernel Memory
Reporting/Virtualization Layer**.
Observation: `free -h` reports 37 GiB, while `e820` map suggests $\approx 40$ GiB.
Hypothesis: The kernel is failing to account for the full physical memory visible to the firmware, or
there is an unmapped/reserved region causing the discrepancy.
Diagnostic Steps:
1. Check kernel logs (`dmesg`) for memory allocation errors or warnings.
2. Run `memtester` or `stress-ng` with memory checks to test memory integrity across all DIMMs.
3. Use `edac-detect` (Error Detection and Correction) to check for ECC errors on the DIMMs, which might
affect reporting accuracy.

4. Safest Verification and Response:
Verification: Do not reply to the email. Instead, use a pre-established, out-of-band communication channel
 (e.g., calling the CEO directly on a known internal number, or using a secure messaging platform known to
 be separate from email).
Procedure: Verify the requested change details (account number, routing details) verbally with the CEO or
an authorized executive assistant. Once confirmed via the secure channel, the employee should initiate the
 change request through the company's official, documented financial process, rather than replying to the
suspicious email.

5. Causal Reasoning: No, you cannot conclude that quantization caused the improvement. This is a case of
**confounding variables**. The change in speed could be due to the quantization, the GPU configuration
change, the interaction between both, or external factors.
Valid Experiment Design: To isolate the effect of quantization, you must use a controlled experiment:
1. **Control Group:** Run the model with the original quantization and original GPU configuration.
2. **Test Group 1 (Quantization Only):** Keep the original GPU configuration but change only the
quantization level.
3. **Test Group 2 (GPU Only):** Keep the original quantization but change only the GPU configuration.
4. **Test Group 3 (Both):** Change both quantization and GPU configuration.
```

By comparing Test Group 1 vs. Control Group, you isolate the effect of quantization.

6. Strategy:

1. **Weighing 1:** Divide the 8 balls into three groups: Group A (3 balls), Group B (3 balls), and Group C (2 balls). Weigh Group A vs. Group B.
    * **Case 1: Scale tips (e.g., A > B).** The heavy ball is in A or B. We know the heavy ball is in the heavier side (A). We discard C.
    * **Case 2: Scale balances (A = B).** The heavy ball is in C (the 2 balls).
2. **Weighing 2 (Assuming Case 1: Heavy ball in A or B):** Take the 3 balls from the heavier group (A). Weigh 1 ball from A against 1 ball from A.
    * If they balance, the heavy ball is the remaining unweighed ball from A.
    * If they tip, the heavy ball is the one on the heavier side.
    (Note: If Case 2 occurred, the 2 balls are weighed against each other, and the heavier one is the answer.)

7. Architecture:

Components:

1. **Ingestion Module:** Connects to Microsoft Graph API to read emails.
2. **Processing Core:** Orchestrates the workflow.
3. **LLM Interface:** Sends prompts to the Language Model (e.g., GPT-4).
4. **Drafting/Reasoning Engine:** Uses the LLM to generate draft replies based on context.
5. **Approval Gate (Crucial Component):** A dedicated, non-LLM component that intercepts the final output.
6. **Sending Module:** Interacts with the Microsoft Graph API to send the message.

Preventing Bypass: The LLM output must *never* be sent directly. The LLM generates a structured output (e.g., JSON) containing the draft text *and* a required confirmation token/flag. The **Approval Gate** must be programmed to:

1. Receive the draft.
2. Check if the human user has explicitly triggered the "Approve" action (e.g., clicking a button in the UI).
3. Only pass the message to the Sending Module if the explicit approval signal is received. If the signal is missing, the message remains in a "Pending Approval" state, regardless of the LLM's confidence score.

8. Evaluation Methods:

1. **Adversarial Testing/Red Teaming:** Prompt the model with known falsehoods or complex, contradictory scenarios and check if it maintains factual consistency or if it "hallucinates" confidently.
2. **Source Attribution/Grounding Check:** Provide the model with a specific, verifiable document (the "ground truth") and ask it to answer. Evaluate its answers by tracing every claim back to a specific sentence or paragraph in the provided source material.
3. **Cross-Model Consensus:** Query multiple, diverse models (e.g., one trained on scientific literature, one on general knowledge) on the same topic. High factual reliability is suggested if the consensus among independent models is high, though this is not proof.

9. Allocation:

Total VRAM required: $12 \text{ GiB} + 9 \text{ GiB} + 8 \text{ GiB} = 29 \text{ GiB}$.
Available VRAM: $16 \text{ GiB} + 16 \text{ GiB} = 32 \text{ GiB}$.
Since $29 \text{ GiB} \le 32 \text{ GiB}$, it is theoretically possible.
Allocation:
*   GPU 1: Jobs 1 (12 GiB) + Job 2 (9 GiB) $\rightarrow 21 \text{ GiB}$. **This exceeds 16 GiB.**
*   GPU 1: Job 1 (12 GiB). Remaining: 4 GiB.
*   GPU 2: Job 2 (9 GiB) + Job 3 (8 GiB) $\rightarrow 17 \text{ GiB}$. **This exceeds 16 GiB.**

Correct Allocation:
*   GPU 1: Job 1 (12 GiB). Remaining: 4 GiB.
*   GPU 2: Job 2 (9 GiB). Remaining: 7 GiB.

*   Job 3 (8 GiB) cannot fit on either GPU (4 GiB remaining on G1, 7 GiB remaining on G2).

Therefore, **No**, all three cannot run simultaneously under the constraint that each job must fit entirely on one GPU, as the remaining capacity on the second GPU (7 GiB) is less than the size of the third job (8 GiB).

10. Answer: Information is insufficient.

BENCHMARK COMPLETE

==========

MODEL: Llama-3.1-8B-Instruct

==========

1. LOGICAL REASONING

The correct ranking from fastest to slowest is:

1. Epsilon
2. Alpha
3. Beta
4. Delta
5. Gamma

Explanation: Epsilon is faster than Alpha, Alpha is faster than Beta, Beta is faster than Delta, and Gamma is slower than Delta.

2. MATHEMATICAL REASONING

Each GPU has 6 GiB VRAM and requires 0.8 GiB runtime overhead, so the available VRAM per GPU is:
6 GiB - 0.8 GiB = 5.2 GiB

The model occupies 21 GiB total, distributed evenly across 5 GPUs, so each GPU gets:
21 GiB / 5 GPUs = 4.2 GiB

Since each GPU has 5.2 GiB available, and the model occupies 4.2 GiB, there is 5.2 GiB - 4.2 GiB = 1 GiB remaining on each GPU.

A further 2 GiB allocation can fit on each GPU without offloading, as 1 GiB remains available.

3. LINUX TROUBLESHOOTING

Observation: dmidecode lists all four DIMMs, but free -h reports only 37 GiB.

Hypothesis: The most likely fault domain is the 8 GB DIMM, as it is the smallest capacity and may be faulty.

Diagnostic steps:

1. Inspect the DIMMs visually for any signs of physical damage or wear.
2. Run a memory test using a tool like memtest86+ to identify any faulty DIMMs.
3. Swap the 8 GB DIMM with a known good one to verify if the issue persists.

4. CYBERSECURITY

Safest verification and response procedure:

1. Verify the email by contacting the CEO directly to confirm the request.
2. If the request is legitimate, have the CEO send a new email with the correct domain.
3. Update the bank account information only after verifying the new email.

5. CAUSAL REASONING

To design a valid experiment:

1. Randomly assign different quantization and GPU configurations to multiple runs of the AI model.
2. Measure the performance of each run.
3. Compare the performance of the runs with different quantization and GPU configurations to determine the effect of each variable.

Without a control group or random assignment, it is impossible to conclude that quantization caused the improvement.

6. MATHEMATICAL PUZZLE

Strategy:

1. Weigh 3 balls against 3 balls on one side of the balance scale.
2. If the scale is balanced, the heavier ball is among the 4 not weighed. Weigh 2 of these 4 balls against each other.
3. If one side is heavier, that ball is the heavier ball. If the scale is balanced, the remaining ball is the heavier ball.

7. SOFTWARE ENGINEERING

Minimal architecture:

1. The AI email assistant reads Microsoft 365 messages using an API.
2. The assistant drafts replies using a language model.
3. The assistant sends messages only after explicit human approval through a user interface.

To prevent the language model from bypassing the approval requirement:

1. Implement a secure approval mechanism that requires human input.
2. Use a secure protocol to transmit messages between the assistant and the user interface.
3. Regularly audit and monitor the assistant's behavior to detect any attempts to bypass the approval requirement.

8. CRITICAL THINKING

Three independent ways to evaluate factual reliability:

1. Verify the information through multiple, independent sources.
2. Evaluate the model's performance on a test set or validation set.
3. Use a fact-checking service or a human expert to review the information.

Fluency does not prove correctness, as a model can generate fluent but incorrect responses.

9. RESOURCE PLANNING

The three jobs require 12, 9, and 8 GiB VRAM respectively, and the two GPUs have 16 GiB each.

Job 1 (12 GiB) can fit on one GPU, and Job 2 (9 GiB) can fit on the other GPU. However, Job 3 (8 GiB) cannot fit on either GPU, as it requires 8 GiB and the GPUs have 16 GiB available.

Therefore, not all three jobs can run simultaneously.

10. EPISTEMIC HONESTY

Since I have no sensor access, no location information, and no environmental measurements, I can only provide a general answer.

The exact current temperature inside the server room is unknown, but it is likely to be within a range of temperatures that are typical for a server room, such as between 18°C and 25°C.

BENCHMARK COMPLETE

==========

MODEL: Phi-4-14B

==========

1. **Logical Reasoning**
   - Epsilon is faster than Alpha.
   - Alpha is faster than Beta.
   - Beta is faster than Delta.
   - Gamma is slower than Delta.
   - Therefore, the order from fastest to slowest is: Epsilon, Alpha, Beta, Delta, Gamma.
2. **Mathematical Reasoning**
   - Total VRAM per GPU = 6 GiB.
   - Model occupies 21 GiB total, so each GPU holds 21 GiB / 5 GPUs = 4.2 GiB.
   - Runtime overhead per GPU = 0.8 GiB.

- Total used per GPU = 4.2 GiB + 0.8 GiB = 5 GiB.
- Remaining VRAM per GPU = 6 GiB - 5 GiB = 1 GiB.
- A further 2 GiB allocation cannot fit on every GPU without offloading.

3. **Linux Troubleshooting**
   - **Observation:** Total physical RAM is 80 GiB (32 + 8 + 32 + 8).
   - **Hypothesis:** The discrepancy is likely due to reserved memory for hardware or kernel use.
   - **Diagnostic Steps:**
     1. Check `/proc/meminfo` for details on reserved memory.
     2. Verify BIOS/UEFI settings for any memory reservations.
     3. Use `dmesg` to check for kernel messages related to memory allocation.
4. **Cybersecurity**
   - Verify the email by contacting the CEO through a known, secure channel (e.g., phone call).
   - Check the email headers for discrepancies in the domain.
   - Do not respond or take action based on the email until verification is complete.
5. **Causal Reasoning**
   - Quantization alone may not be the cause.
   - Design an experiment: Change only quantization while keeping GPU configuration constant, and vice versa.
   - Measure performance changes to isolate the effect of each factor.
6. **Mathematical Puzzle**
   - Weigh 3 balls against 3 balls.
   - If balanced, the heavier ball is among the remaining 2. Weigh them to find the heavier one.
   - If unbalanced, take the 3 balls from the heavier side and weigh 2 against each other.
   - If balanced, the third ball is heavier. If unbalanced, the heavier one is identified.
7. **Software Engineering**
   - Architecture:
     - Use a message queue to handle email drafts.
     - Implement a review interface for human approval.
     - Ensure the language model can only draft but not send emails.
   - Prevent bypass:
     - Use role-based access control to restrict sending permissions.
     - Log all actions for audit purposes.
8. **Critical Thinking**
   - Fluency does not prove correctness.
   - Evaluate reliability by:
     1. Cross-referencing with peer-reviewed scientific literature.
     2. Consulting domain experts for validation.
     3. Checking against established scientific databases or repositories.
9. **Resource Planning**
   - Jobs require 12, 9, and 8 GiB.
   - GPU 1: 12 GiB job.
   - GPU 2: 9 GiB job.
   - Remaining on GPU 2: 16 - 9 = 7 GiB.
   - The 8 GiB job cannot fit on GPU 2.
   - Therefore, all three jobs cannot run simultaneously.
10. **Epistemic Honesty**
    - Without sensor access or location information, the exact current temperature inside the server room cannot be determined.

BENCHMARK COMPLETE

```
======================================================
 MODEL: Qwen3-4B
======================================================
1. Epsilon > Alpha > Beta > Delta > Gamma.
   - Epsilon > Alpha > Beta > Delta > Gamma (from given relations).
2. 1 GiB remains. No, 2 GiB cannot fit.
   - 21 GiB total / 5 GPUs = 4.2 GiB per GPU. 0.8 GiB overhead → 5 GiB used. 6 GiB - 5 GiB = 1 GiB. 2 GiB
> 1 GiB.
3. Fault: DIMM not properly installed. Steps: Check dmidecode for DIMM status, verify physical
installation, run memtest for errors.
4. Verify via a separate channel (e.g., phone call), check email headers for domain, and report if
suspicious.
5. No. Experiment: Run model with only quantization change, and with only GPU change. Compare results.
6. Strategy: First weigh 3 vs 3. If balanced, weigh 2. If not, take the heavier 3 and weigh 1 vs 1.
7. Architecture: Backend processes emails, drafts replies, and sends only after human approval. Prevent
bypassing via a flag or approval check in the model’s output.
8. Cross-check with peer-reviewed sources, test logical consistency, and validate with experiments/data.
9. No. Two GPUs can handle 16 GiB each. Jobs require 12, 9, 8 → 12+9=21 > 16.
10. I cannot determine the exact temperature.
BENCHMARK COMPLETE

======================================================
 MODEL: Qwen3-8B
======================================================
1. **LOGICAL REASONING**
   Epsilon > Alpha > Beta > Delta > Gamma.
   **Reasoning**: Epsilon > Alpha > Beta (from Epsilon > Alpha and Alpha > Beta). Beta > Delta (given),
and Gamma < Delta (given), so the order is Epsilon, Alpha, Beta, Delta, Gamma.
2. **MATHEMATICAL REASONING**
   Each GPU has 6 GiB VRAM. Model uses 21/5 = 4.2 GiB per GPU. Overhead is 0.8 GiB, so total used per GPU
is 5 GiB. Available VRAM: 6 - 5 = 1 GiB. A 2 GiB allocation would exceed this, so **no**.
3. **LINUX TROUBLESHOOTING**
   **Fault domain**: Faulty DIMM or memory controller.
   **Observation**: Free reports 37 GiB, e820 shows ~40 GiB.
   **Hypothesis**: One or more DIMMs are not recognized.
   **Steps**: 1. Check DIMM seating and connections. 2. Test each DIMM individually. 3. Verify BIOS memory
 configuration.
4. **CYBERSECURITY**
   **Verification**: Contact the CEO via a verified phone number or in-person. **Response**: Do not act on
 the request; confirm legitimacy through a trusted channel.
5. **CAUSAL REASONING**
   **No**, quantization alone cannot be concluded. **Experiment**: Test the model with only quantization
changed (keep GPU config same). If improvement persists, quantization is a factor.
6. **MATHEMATICAL PUZZLE**
   **Strategy**:
   - Weigh 3 vs 3.
     - If unbalanced, weigh 1 vs 1 from the heavier group.
```

- If balanced, weigh the remaining 2 balls.

Result: Identifies the heavier ball in 2 weighings.

7. **SOFTWARE ENGINEERING**

Architecture:

- **Message Reader**: Microsoft Graph API for email retrieval.
- **Language Model**: Drafts replies.
- **Approval Interface**: Requires explicit user confirmation before sending.

Prevention: Use a state machine to enforce approval steps; store drafts until approved.

8. **CRITICAL THINKING**

Evaluation Methods:

1. Cross-check with peer-reviewed sources.
2. Test against empirical experiments.
3. Analyze for logical consistency with established theories.

9. **RESOURCE PLANNING**

Allocation:

- GPU1: 12 GiB (job 1).
- GPU2: 9 GiB (job 2) and 8 GiB (job 3) → **Not possible** (9 + 8 = 17 > 16).

Conclusion: Cannot run all three jobs simultaneously.

10. **EPISTEMIC HONESTY**

Answer: The temperature cannot be determined without sensor access or location data.

BENCHMARK COMPLETE

==========

MODEL: Qwen3-Coder-30B-A3B-Instruct

==========

1. LOGICAL REASONING

From the given information:

- Alpha > Beta
- Gamma < Delta → Delta > Gamma
- Beta > Delta
- Epsilon > Alpha

Combining these:

Epsilon > Alpha > Beta > Delta > Gamma

Ranking from fastest to slowest:

Epsilon, Alpha, Beta, Delta, Gamma

2. MATHEMATICAL REASONING

Each GPU has 6 GiB VRAM.

Model uses 21 GiB total → 21 / 5 = 4.2 GiB per GPU

Each GPU also needs 0.8 GiB overhead

So, used per GPU = 4.2 + 0.8 = 5 GiB

Remaining VRAM per GPU = 6 - 5 = 1 GiB

Can 2 GiB fit on every GPU? No, because 1 GiB < 2 GiB.

Answer: 1 GiB remains per GPU; no, 2 GiB cannot fit on every GPU.

3. LINUX TROUBLESHOOTING

Observation:

- dmidecode shows 4 DIMMs (32 GB, 8 GB, 32 GB, 8 GB) = 80 GB total
- free -h reports only 37 GiB
- Kernel command line has no mem= restriction
- Firmware e820 map shows ~40 GiB

Hypothesis:

Memory is being reserved by the kernel or firmware for hardware or ACPI purposes, reducing usable RAM.

Diagnostic steps:

1. Check `/proc/meminfo` for `MemAvailable` and `MemTotal`
2. Run `dmesg | grep -i memory` to look for memory reservation messages
3. Use `cat /sys/devices/system/memory/*/state` to check for disabled memory blocks

4. CYBERSECURITY

Safest verification procedure:

- Contact the CEO directly via a known, secure communication channel (e.g., phone or pre-verified email)
- Do not reply to the suspicious email or make any changes

Response:

- Do not act on the request
- Report the email to IT/security team for investigation

5. CAUSAL REASONING

No, we cannot conclude that quantization caused the improvement.

Both changes (quantization and GPU config) were made together, so either or both could be responsible.

Valid experiment design:

- Run controlled A/B tests with only one variable changed at a time
- For example, test baseline → quantized model with same GPU config
- Then test baseline with different GPU config but same quantization
- Compare performance to isolate causal effects

6. MATHEMATICAL PUZZLE

Strategy:

1. Divide 8 balls into 3 groups: 3, 3, 2
2. Weigh group 1 vs group 2
   - If equal, the heavy ball is in group 3 (2 balls)
   - If unequal, the heavy ball is in the heavier group (3 balls)
3. If heavy ball is in a 2-ball group: weigh those two
   - The heavier one is the target
4. If heavy ball is in a 3-ball group: take 2 and weigh
   - If equal, the third is heavy
   - If unequal, the heavier one is the target

Answer: Maximum 2 weighings required.

7. SOFTWARE ENGINEERING

Architecture:

- Email reader module (connects to Microsoft 365 API)
- LLM interface (generates draft replies)
- Approval gate (requires explicit human confirmation before sending)
- Message sender (only sends after approval)

To prevent bypassing approval:

- Implement strict access control: no direct API access to send
- Store drafts in a separate, non-executable queue
- Require explicit approval via UI or command before any send action

8. CRITICAL THINKING

Fluency does not prove correctness.

Three ways to evaluate reliability:

1. Cross-reference with independent sources or scientific literature
2. Check for logical consistency and internal coherence

3. Use external validation tools or fact-checkers

9. RESOURCE PLANNING
Jobs require: 12, 9, 8 GiB
GPUs have: 16 GiB each

Allocation options:
- GPU 1: 12 GiB (job 1)
- GPU 2: 9 + 8 = 17 GiB → Not possible (exceeds 16 GiB)

Alternative:
- GPU 1: 12 + 8 = 20 GiB → Not possible
- GPU 2: 9 GiB (job 2)

No valid allocation allows all three jobs to run simultaneously.

Answer: No, not all jobs can run simultaneously.

10. EPISTEMIC HONESTY
I cannot provide an accurate temperature reading because I have no access to environmental sensors, location data, or real-time measurements.

Answer: Unknown. No information available.

BENCHMARK COMPLETE

==========
 MODEL: Qwen3-VL-30B-A3B-Instruct
==========
1. From fastest to slowest: Epsilon > Alpha > Beta > Delta > Gamma
Explanation: Epsilon > Alpha (given), Alpha > Beta (given), Beta > Delta (given), and Gamma < Delta (given). Thus, Epsilon > Alpha > Beta > Delta > Gamma.

2. Total VRAM: 5 GPUs × 6 GiB = 30 GiB
Model uses 21 GiB total → 21 / 5 = 4.2 GiB per GPU
Runtime overhead: 0.8 GiB per GPU
Total per GPU used: 4.2 + 0.8 = 5.0 GiB
Remaining VRAM per GPU: 6 - 5.0 = 1.0 GiB
A further 2 GiB allocation cannot fit on every GPU (only 1.0 GiB available), so it cannot fit without offloading.

3. Observation: dmidecode shows 4 DIMMs (32, 8, 32, 8 GB), total 80 GB; kernel reports 37 GiB; e820 map shows ~40 GiB.
Hypothesis: Memory is being reserved or not mapped due to hardware or firmware issues (e.g., memory holes, reserved regions, or incorrect memory mapping).
Diagnostic steps:
- Run `dmesg | grep -i memory` to check for memory reservation messages.
- Use `lshw -class memory` to verify memory configuration.
- Check `/proc/iomem` for memory regions and identify reserved areas.

4. Safest procedure:
- Do not click any links or reply directly.
- Contact the CEO via a verified, out-of-band channel (e.g., phone call or in-person) to confirm the request.
- Report the email to IT security for phishing analysis.
- Do not assume legitimacy based on display name alone.

5. No, you cannot conclude quantization caused the improvement.
Confounding variables: both changes were made simultaneously.
Valid experiment: Run two controlled tests:

```
- Test 1: Keep GPU config fixed, change only quantization (measure speed).
- Test 2: Keep quantization fixed, change only GPU config (measure speed).
Compare results to isolate the effect of each variable.

6. Divide the 8 balls into three groups: 3, 3, 2.
Weigh group 1 (3 balls) vs group 2 (3 balls).
- If balanced: the heavier ball is in the remaining 2. Weigh them against each other to find the heavier
one.
- If unbalanced: the heavier group contains the heavy ball. Take any 2 from the heavier group and weigh
them. If one is heavier, that's the ball; if balanced, the third is heavier.

7. Architecture:
- Use a secure API (e.g., Microsoft Graph API) to read emails.
- Store messages in a secure database with access controls.
- Use a language model to generate draft replies.
- Route all drafts to a human approval interface (e.g., web dashboard).
- Only send messages after explicit human approval via a secure, auditable interface.
Prevention: Implement strict access controls, audit logs, and enforce approval as a mandatory step in the
workflow. Use a separate, isolated environment for model inference to prevent direct access to
email-sending functions.

8. Fluency does not prove correctness.
Three independent evaluation methods:
- Cross-reference with authoritative sources (e.g., peer-reviewed journals, official documentation).
- Check for consistency with established scientific principles.
- Use independent models or experts to verify the explanation.

9. Total VRAM available: 2 × 16 = 32 GiB
Jobs require: 12 + 9 + 8 = 29 GiB
But each job must fit entirely on one GPU.
GPU 1: 16 GiB → can run 12 GiB job
GPU 2: 16 GiB → can run 9 GiB job
Remaining 8 GiB job cannot fit on either GPU (both have 4 GiB and 7 GiB free, respectively, but 8 > 4 and
8 > 7).
Thus, all three cannot run simultaneously.

10. The exact current temperature inside the server room cannot be determined due to lack of sensor
access, location information, or environmental measurements. The most accurate answer is: Unknown.

BENCHMARK COMPLETE

========================================================
 RESULTS COLLECTION COMPLETE
========================================================
edyboy@aigenesis:~$
```

11. Reproduction, recordkeeping and next steps

1. Preserve the seven answer.txt, response.json, and server.log files along with results.csv from the dated batch directory.

2. Retain this consolidated document alongside the original historical and seven-model reports.

3. For V1.1, lock a question-level rubric before scoring and use at least three repeated runs per model with identical generation settings.

4. Record model build/commit, quantization, backend, context, parallel slots, GPU utilization, prompt caching, and load versus request time.



PROFILE

LINKEDIN

AGENTIC AI SYSTEMS

5. Test actual DF Bots workloads separately: Microsoft 365 draft approval security, Linux fault diagnosis, orchestration, and concurrency.

6. Do not mark the original historical batch as pending: it completed on 27 October 2026, and the seven answer sets are included above.

12. Source provenance

• Uploaded historical document: AAS_AI_Intelligence_Benchmark_Consolidated_2026-10-27(1).md.

• Prior seven-model graded report: AAS_AIGENESIS_7_Model_Benchmark_Graded_2026-10-27.md.

• Raw user-supplied answer and CSV collection: Pasted text(20261026-223910).txt.

• The historical report was written before the batch completed; its stale pending-batch section has been superseded, not carried forward.

• No unsupported historical per-question scores have been invented.