

Trustworthy Compute Characteristics of Educational AI Workloads



From generating examination questions to measuring a dependable, auditable workload

The problem: generation is not deployment

An examination system must be dependable, measurable and accountable—not merely fluent.

01

Operational uncertainty

How long will a batch take? How much CPU and memory will it consume?

02

Incomplete delivery

How many requested questions become saved drafts- and what happens when generation fails?

03

Governance gap

Can we reconstruct the settings, trace the run and link outputs to human review?

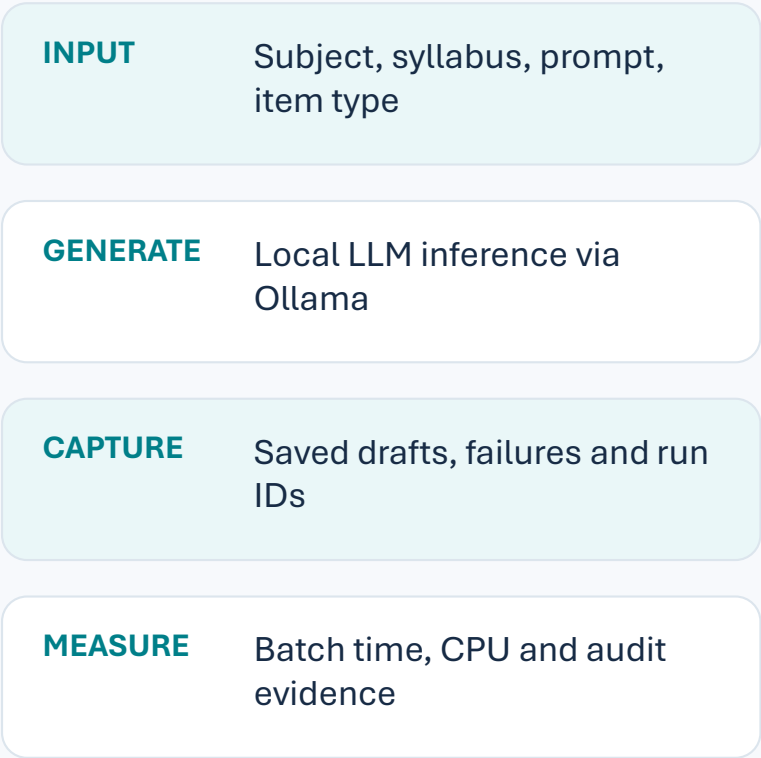
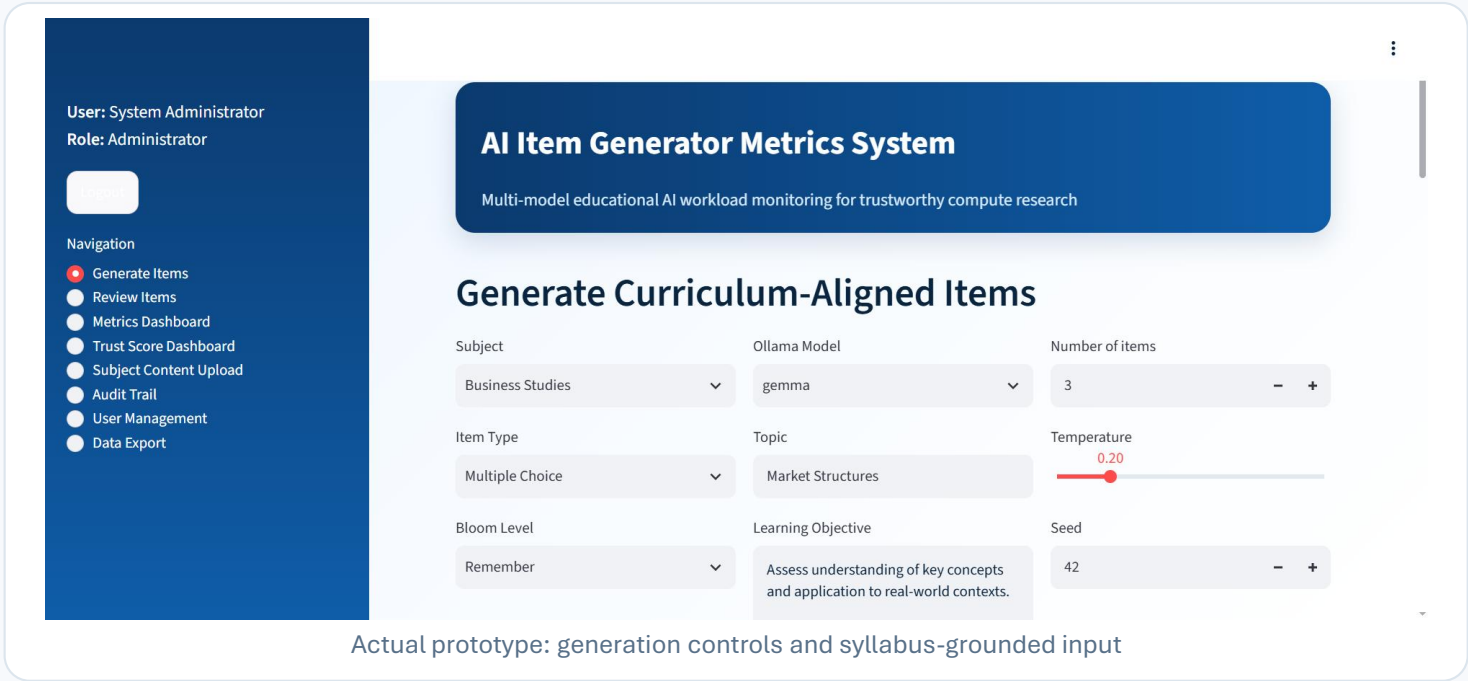
Research question: What makes locally hosted educational AI a trustworthy compute workload?

Research objectives: from output to evidence

- | | | |
|----------|------------------------------------|--|
| 1 | Characterise workload | Record batch duration, CPU use, memory evidence and output throughput. |
| 2 | Measure operational yield | Compare requested versus saved items and document failed attempts. |
| 3 | Compare models and subjects | Describe observed patterns across four models and four core subjects. |
| 4 | Examine operational trust | Inspect hashes, audit events, review activity and governance evidence. |
| 5 | Specify the next benchmark | Design controlled replay, human-review linkage and infrastructure profiling. |

The prototype: a local, instrumented AI workload

Ollama-hosted Gemma, Llama 3, Mistral and Phi; item generation is logged and measured.



Deployment described in study: HP Windows server, 64 GB RAM. CPU observations are server-wide, not model-only.

The Controls used

.

Trustworthy compute

Compute behaviour that is measurable, repeatable under specified conditions, traceable and governable.

Temperature

A sampling parameter: lower values generally reduce variation; higher values allow more diverse responses.

Seed

A specified random starting state that helps repeat a sampling sequence when other conditions match.

Reproducibility

The ability to replay documented conditions and test whether outputs and performance repeat.

Methodology: an observational compute study

A recorded generation run is the unit of analysis; the current cohort is not a matched benchmark.

01

SELECT

- Subject
- model
- topic
- requested count

02

RUN

- Local inference;
- generated draft text

03

LOG

- Run ID
- saved items
- failure status

04

PROFILE

- Elapsed batch time
- server CPU
- other metrics

05

ANALYSE

- Yield
- descriptive contrasts
- evidence gaps

Scope: Geography • History • Business Studies • Economics

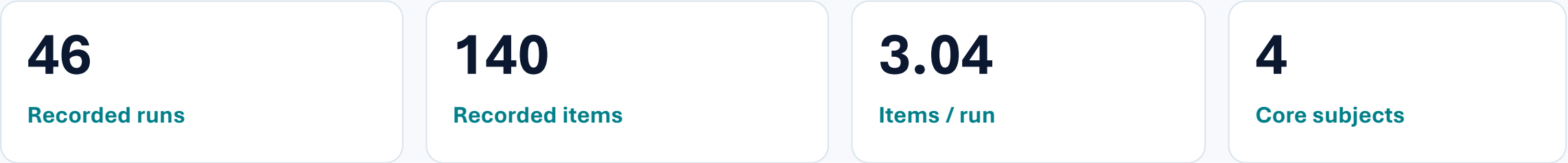
2025: preliminary educational-quality study. 2026: compute and operational evidence; new SME results pending.

Metric dictionary: what the numbers mean

Batch duration	Wall-clock time from start to end of a batch	seconds
Average / peak CPU	Mean and maximum sampled server processor use	% of server
Throughput	Saved items ÷ batch duration × 60	saved items/min
Operational save rate	Saved items ÷ requested items × 100	%
Memory	Recorded or peak memory footprint, where available	MB
Latency	Usually response delay; prototype field stores total batch time	not per-item latency

Results: the reconciled four-subject cohort

46 recorded runs produced 140 saved or recorded drafts.



SUBJECT	RUNS	RECORDED ITEMS	
Geography	17		49
History	12		38
Business Studies	13		33
Economics	4		20

Fairer comparison: equal output, not equal experiments

Estimated run equivalents to obtain 20 recorded items at each subject’s historical yield.

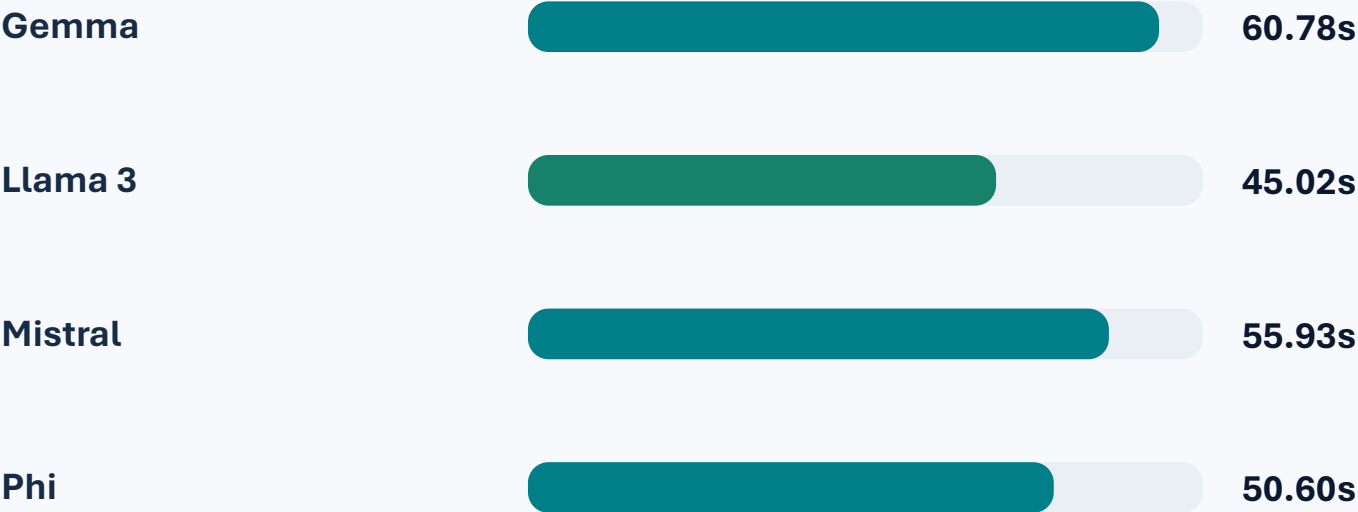


Formula: $20 \times \text{observed runs} \div \text{observed items}$ • Fractional run equivalents are estimates, not new tests.

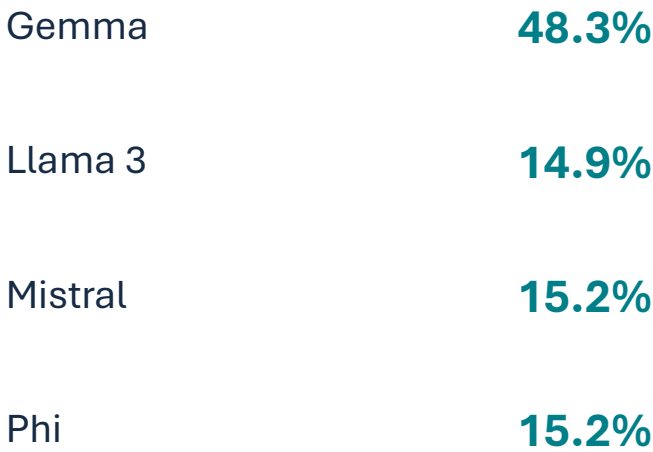
Observed model contrast: Economics batches

Four model runs each saved 5/5 items; duration and CPU still differed.

BATCH DURATION (seconds)



AVG SERVER CPU



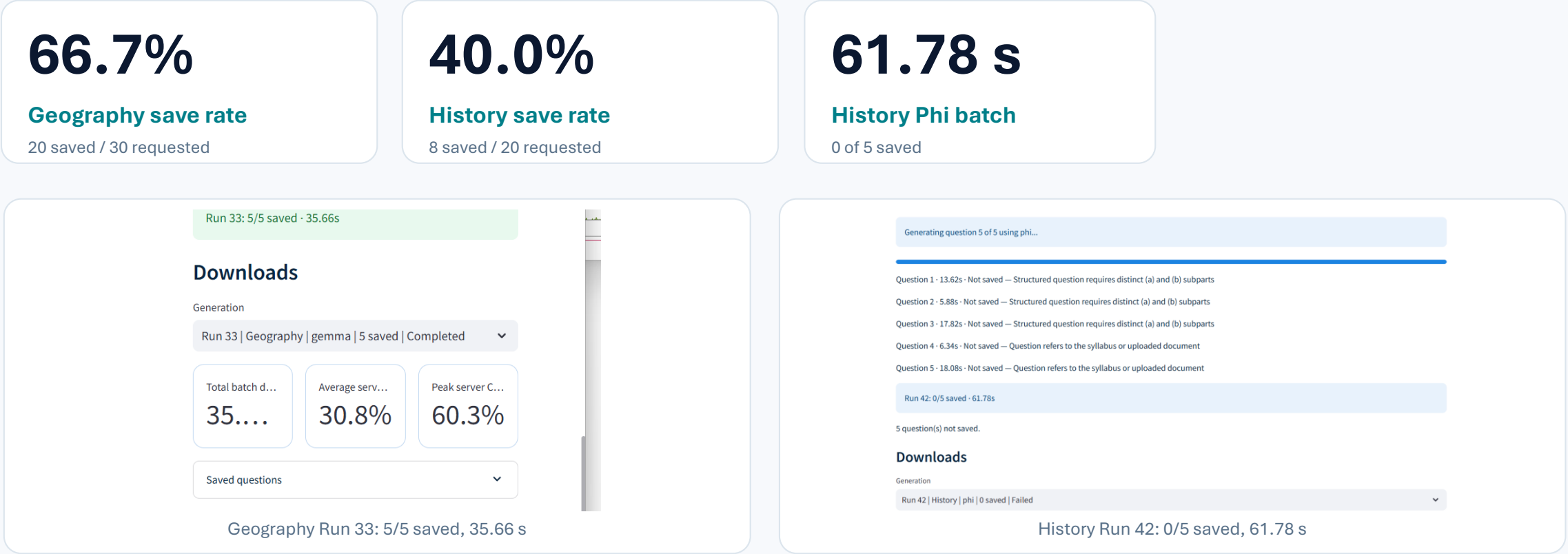
One model, four subjects: different workload profiles

Selected Gemma runs illustrate why yield, time and CPU must be read together.

SUBJECT / RUN	SAVED	TIME (s)	AVG CPU	PEAK CPU	ITEMS/min
Geography · 33	5/5	35.66	30.8%	60.3%	8.41
History · 39	1/5	48.92	44.6%	76.8%	1.23
Economics · 49	5/5	60.78	48.3%	80.5%	4.94
Business · 53	5/5	53.80	45.9%	65.8%	5.58

Example: History Run 39 consumed 48.92 s but saved only one item.

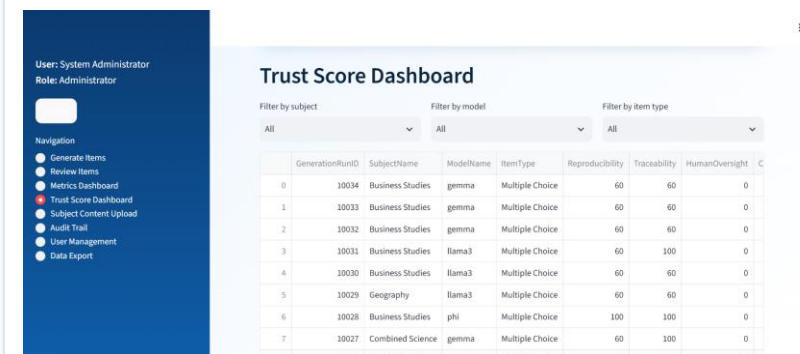
Reliability finding: failed output still costs compute



Operational trust: evidence

The prototype combines record presence and review activity into a provisional index.

R	Run hash exists	Reproducibility evidence
T	Audit event exists	Traceability evidence
H	Review decisions / saved items	Human-review coverage
C	Compute-metrics exists	Compute evidence
S	Grounding record exists	Context evidence



User: System Administrator
Role: Administrator

Navigation

- Generate Items
- Review Items
- Metrics Dashboard
- Trust Score Dashboard
- Subject Content Upload
- Audit Trail
- User Management
- Data Export

Trust Score Dashboard

Filter by subject: All | Filter by model: All | Filter by item type: All

	GenerationRunID	SubjectName	ModelName	ItemType	Reproducibility	Traceability	HumanOversight	C
0	10034	Business Studies	gemma	Multiple Choice	60	60	0	
1	10033	Business Studies	gemma	Multiple Choice	60	60	0	
2	10032	Business Studies	gemma	Multiple Choice	60	60	0	
3	10031	Business Studies	llama3	Multiple Choice	60	100	0	
4	10030	Business Studies	llama3	Multiple Choice	60	60	0	
5	10029	Geography	llama3	Multiple Choice	60	60	0	
6	10028	Business Studies	phi	Multiple Choice	100	100	0	
7	10027	Combined Science	gemma	Multiple Choice	60	100	0	

Actual trust dashboard

$$\text{Governance} = (R + T + H + S) / 4$$

$$\text{Overall} = (R + T + H + C + \text{Governance}) / 5$$

Trust- Traceability through Audit Trail

	AuditLogID	EntityName	EntityID	ActionName	ActionBy	ActionDetails
8	275	GeneratedItem	206	Saved sequential item	admin	Run=66; Item=4; Type=Multiple Choice; DurationSeconds=3.62; Attempts=1; Rev
9	274	GeneratedItem	205	Saved sequential item	admin	Run=66; Item=3; Type=Multiple Choice; DurationSeconds=3.61; Attempts=1; Rev
10	273	GeneratedItem	204	Saved sequential item	admin	Run=66; Item=2; Type=Multiple Choice; DurationSeconds=3.52; Attempts=1; Rev
11	272	GeneratedItem	203	Saved sequential item	admin	Run=66; Item=1; Type=Multiple Choice; DurationSeconds=7.87; Attempts=1; Rev
12	271	GenerationRun	65	Sequential generation finished	admin	Status=Completed; Saved=5/5; DurationSeconds=22.14; GroundingDocs=2; Cpu
13	270	GeneratedItem	202	Saved sequential item	admin	Run=65; Item=5; Type=Multiple Choice; DurationSeconds=4.19; Attempts=1; Rev
14	269	GeneratedItem	201	Saved sequential item	admin	Run=65; Item=4; Type=Multiple Choice; DurationSeconds=4.22; Attempts=1; Rev
15	268	GeneratedItem	200	Saved sequential item	admin	Run=65; Item=3; Type=Multiple Choice; DurationSeconds=4.20; Attempts=1; Rev
16	267	GeneratedItem	199	Saved sequential item	admin	Run=65; Item=2; Type=Multiple Choice; DurationSeconds=4.27; Attempts=1; Rev
17	266	GeneratedItem	198	Saved sequential item	admin	Run=65; Item=1; Type=Multiple Choice; DurationSeconds=5.11; Attempts=1; Rev
18	265	GenerationRun	64	Sequential generation finished	admin	Status=Completed; Saved=5/5; DurationSeconds=27.17; GroundingDocs=2; Cpu

Trust-Quality control of output through a Human in the loop

<<

User: System Administrator
Role: Administrator

Logout

Navigation

- Generate Items
- Review Items
- Metrics Dashboard
- Trust Score Dashboard
- Subject Content Upload
- Audit Trail
- User Management
- Data Export

0000-00-00

Reviewer Quality Score

80

Reviewer Clarity Score

4.00

Reviewer Alignment Score

90

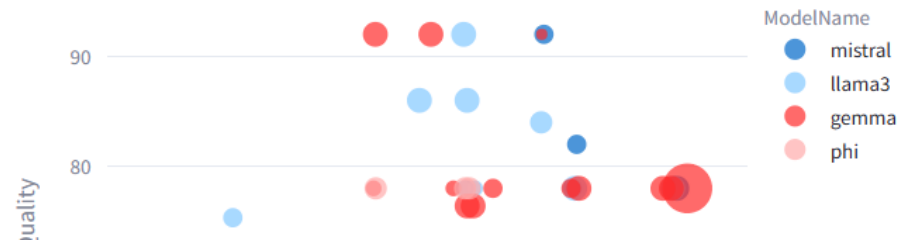
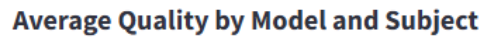
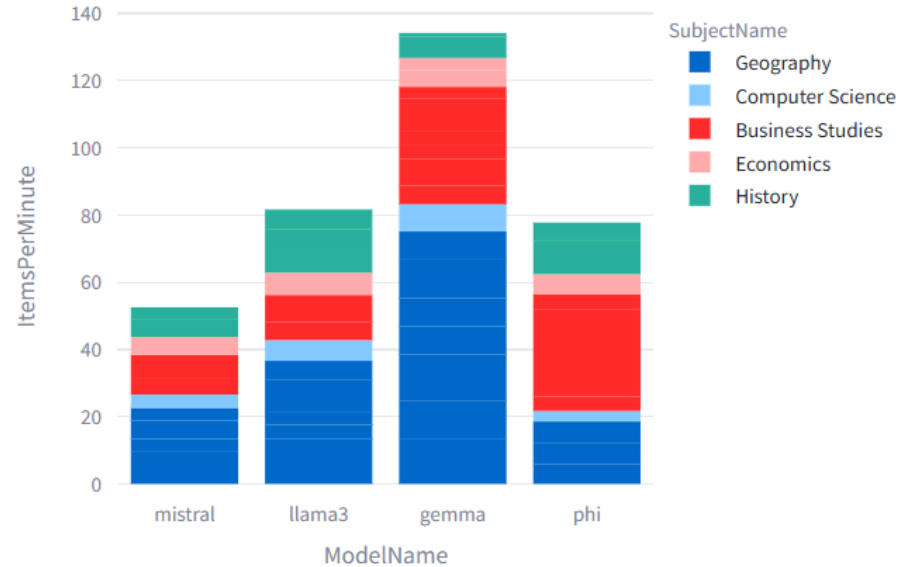
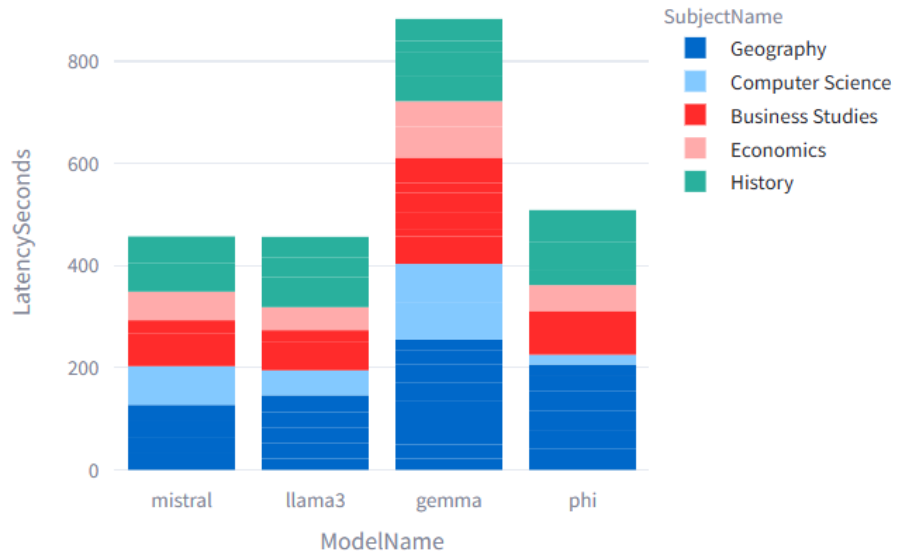
Review Comment

Save Review

ZIMSEC | Dr Louisa Muparuri | Trustworthy compute

04

Trust- Compute observability evidence



Trust- context evidence

Upload Syllabi and Subject-Specific Context

Subject

Computer Science

Content type

Syllabus

Upload PDF, Word, text, CSV or image files

 Drag and drop files here
Limit 200MB per file • PDF, DOCX, TXT, CSV, MD, PNG, JPG, JPEG

Browse files

Save uploaded content

Existing uploaded context

	SubjectContentID	ContentTitle	ContentType	OriginalFileName	UploadedBy	UploadedAt	TextLength
0	9	computer science ordinary level	Specification Gr	computer science ordinary level.docx	admin	2026-10-13 13:59:53	162
1	8	CamScanner 18-02-2026 10.08	Past Paper	CamScanner 18-02-2026 10.08.pdf	admin	2026-10-13 13:57:17	71
2	7	2025 Olevel P2	Past Paper	2025 Olevel P2.pdf	admin	2026-10-13 13:57:17	14
3	6	Computer Science O level Syllabus	Syllabus	Computer Science O level Syllabus.pdf	admin	2026-10-13 13:54:55	61,378

Where are we in this research?

Multi-model, multi-subject study

ACHIEVED

4 models; 4 core subjects; 46 runs / 140 items

Compute characteristics

PARTIAL

Batch duration, CPU and yield measured; RAM and per-item latency incomplete

Reproducibility and audit completeness

PARTIAL

Hashes and audit features exist; replay and completeness not verified

Reviewer agreement and acceptance

PENDING

2026 SME returns not complete by reporting cutoff

Educational AI Trust Framework

PROPOSED

Operational index exists; weighting and validity not established

Way forward: reproducibility by design

Move from an observable pilot to a controlled, replayable HPC benchmark.

1

PIN

Record model digest, prompt and syllabus versions

2

CONTROL

Apply and log seed, temperature and token settings

3

REPEAT

5 repeats $\times$ 4 models $\times$ 4+ subjects = 80+ runs

4

PROFILE

Process CPU, peak RAM, tokens/s and request latency

5

VALIDATE

Replay outputs, audit completeness and SME review

Design controls first $\rightarrow$ verify reproducibility through repeated measurements.

Infrastructure opportunity: from pilot to HPC research

Exploring the potential repurposing of a decommissioned DIRAC cluster for the next research phase.

WHY HPC CAPACITY MATTERS

- **Controlled comparisons** Repeat the same workloads across models, subjects and seeds.
- **Better compute profiling** Measure CPU, memory, throughput, failures and energy under known conditions.
- **Governed research environment** Support local deployment, audit records and human-review workflows.

A responsible pathway

- 1 Technical feasibility
- 2 Institutional agreement
- 3 Security & operating costs
- 4 **Pilot HPC benchmark**

Collaboration invitation | Cluster assessment, reproducible benchmarking and sustainable research operations.

The DIRAC cluster is an opportunity under exploration-WIP.

THE RESEARCH CONTRIBUTION

From generating questions to engineering trust.

DEMONSTRATED

A working, instrumented local AI item generator

MEASURED

Output yield, batch time, server CPU and failures

NEXT

Controlled reproducibility, SME validation and HPC readiness

The opportunity: reproducible, auditable educational AI research for national assessment.