

Benchmarking Open-Weight vs. Frontier LLMs on African Health and Financial-Inclusion Reasoning, With and Without Graph RAG

1. Impact Potential: AI in African Healthcare and Financial Services

AI is poised to reshape two of Africa's highest-stakes domains. In healthcare, the continent carries roughly 25% of the global disease burden with only about 3% of the world's healthcare professionals, and faces a projected shortfall of 4.3 million doctors by 2035 (Brookings Institution, 2025). The African Development Bank projects AI-driven productivity gains concentrated in health and life sciences alongside agriculture and finance, while initiatives such as the Gates Foundation–OpenAI Horizon1000 partnership (\$50M) explicitly target AI-assisted triage, clinical decision support, and multilingual communication to offset workforce scarcity (Medical Device Network, 2026; African Development Bank, 2025).

In financial services, MSMEs constitute 80–90% of African employment yet face a formal financing gap exceeding \$5.2 trillion in emerging markets, with Sub-Saharan Africa showing the slowest growth in MSME finance supply of any region (International Finance Corporation, 2025; SME Finance Forum, 2025). AI-driven credit scoring, mobile-money-linked underwriting, and microinsurance distribution are positioned as the most plausible near-term levers to close this gap, particularly as FinScope MSME survey data shows financial inclusion increasingly mediated by mobile money rather than traditional banking infrastructure (FinMark Trust, 2022; Reserve Bank of Zimbabwe, 2022).

2. Motivation: Compute Scarcity and Data-Colonial Dependency

Both domains illustrate a structural vulnerability: African AI deployment is overwhelmingly dependent on foreign-hosted frontier models, foreign cloud compute, and a semiconductor supply chain in which the continent has negligible manufacturing capacity. This mirrors a "data-colonial" pattern in which value-bearing data (clinical, financial, behavioral) flows outward to be processed on infrastructure the continent does not control, returning as a paid service rather than sovereign capability. Heavy reliance on frontier-model APIs compounds this: it ties critical health and financial inference pipelines to external pricing, rate limits, content policies,

and geopolitical export-control decisions outside African institutions' control. Building locally-deployable, open-weight pipelines — even at reduced parameter counts — is therefore not merely a cost question but a sovereignty question: can African research institutions retain meaningful control over the AI systems mediating health and financial decisions affecting their populations?

Research question. Can lightweight open-source models, grounded with strongly aligned Graph RAG, achieve results comparable to frontier models in Africa's most critical domains — reducing dependence on heavy compute, cost, and foreign hardware?

3. Datasets

Health — AfriMed-QA. A Pan-African, multi-specialty medical question-answering benchmark comprising 15,000 questions sourced from over 60 medical schools across 16 African countries and 32 specialties, designed to evaluate LLMs on correctness and demographic bias in African clinical contexts (Olatunji et al., 2024). For this experiment, a stratified random sample was drawn across specialties to keep frontier-API and local-inference costs tractable.

Financial — DataOrg Financial Health Prediction Challenge (Zindi). A FinScope MSME survey instrument administered by FinMark Trust across Eswatini, Zimbabwe, Malawi, and Lesotho, classifying business owners into Low/Medium/High financial health based on demographic, cash-flow, credit, and insurance-access features (FinMark Trust, 2022, 2024). Because the raw target distribution is heavily imbalanced (~65% Low/30% Medium/5% High), the benchmark sample was balanced equally across classes so accuracy reflects genuine discrimination rather than majority-class guessing.

4. Methodology

Graph RAG construction. Two knowledge bases were built using a shared pipeline: PDF/markdown ingestion → chapter/section segmentation → cleaning → overlapping chunking (~300–350 words) → entity extraction (heuristic/NER) → graph construction (`mentioned_in`, `co_occurs_with`, `part_of` edges) → export to a chunk store plus graph (GraphML/JSON). The health KB was built from *Current Medical Diagnosis & Treatment 2025*, with chapter segmentation auto-parsed from the book's own two-column table of contents (44 chapters) and parallelized page extraction (a single-PDF-reopen bug fix yielded a ~7.4× speedup, compounding further with process-pool parallelism). The financial KB reused the project's own MSME financial-health markdown document, whose explicit "Entity Relationship" bullets were parsed directly into graph edges — making it inherently more graph-structured than the health corpus.

Retrieval. A hybrid `GraphRAGRetriever` embeds the query and chunk texts (`all-MiniLM-L6-v2`), retrieves top-k chunks by cosine similarity, then expands one hop via

shared entities in the graph — surfacing chunks that share a strongly co-occurring entity even without lexical overlap with the query.

Model conditions. Six conditions per dataset: two frontier baselines (Gemini Pro, GPT-5) run without RAG, representing normal frontier operating conditions; and two open-weight models — Qwen3.6-27B (via OpenRouter, reasoning-enabled) and Gemma-4-31B-it (via the Gemini API) — each run with and without Graph RAG context injected into the prompt. This substitution for genuinely local GGUF inference was necessitated by GPU constraints (see Limitations).

Grading. A Gemini Judge (temperature 0) scored each response 1 (correct) or 0 (incorrect) against the reference answer, explicitly instructed to ignore formatting, verbosity, and presentation differences and judge only semantic correctness — controlling for the fact that models vary widely in answer formatting conventions.

5. Results

Health (AfriMed-QA), N≈32 sampled questions:

Condition	Accuracy
GPT-5	0.875
Qwen3.6-27B (no RAG)	0.813
Qwen3.6-27B + RAG	0.719
Gemma-4-31B-it + RAG	0.656
Gemma-4-31B-it (no RAG)	0.594
Gemini Pro	0.188

Financial (DataOrg MSME), N≈60 sampled rows:

Condition	Accuracy
GPT-5	0.517
Qwen3.6-27B (no RAG)	0.400
Qwen3.6-27B + RAG	0.333
Gemma-4-31B-it (no RAG)	0.083
Gemma-4-31B-it + RAG	0.033

Gemini Pro

0.000

Patterns. GPT-5 led both domains, consistent with frontier-model expectations. More strikingly, Qwen3.6-27B without RAG outperformed every other non-GPT-5 condition in both domains, including its own RAG-augmented variant — a counterintuitive result suggesting the injected context may have introduced noise, distracting reasoning, or conflicted with the model's parametric knowledge rather than reinforcing it. Gemma showed the opposite pattern in health (RAG improved accuracy, 0.594→0.656) but the same RAG-degradation pattern in finance (0.083→0.033), suggesting RAG benefit is domain- and retrieval-quality-dependent rather than uniformly positive. Gemini Pro's near-zero scores across both domains are anomalous relative to its known general capability and most plausibly reflect an artifact of this pipeline — e.g., API response formatting, safety-filter truncation, or judge/extraction mismatch — rather than genuine reasoning failure, and should not be read as a real capability finding without further diagnostic work.

6. Conclusions

The evidence is mixed on the research question. Qwen3.6-27B without RAG came within ~6–12 accuracy points of GPT-5 in health and financial reasoning respectively, suggesting a 27B open-weight model accessed via API can approach frontier performance on these domains without local heavy compute — a meaningful sovereignty-relevant result on its own, since it reduces (though does not eliminate, given OpenRouter/API dependency) hardware lock-in. However, the central hypothesis — that *Graph RAG* specifically would let lightweight models close the gap with frontier models — was **not** supported here: RAG degraded performance in three of four open-weight conditions. This points to retrieval-quality and prompt-integration issues (chunk relevance, context length competing with reasoning, entity-graph noise) as the binding constraint rather than the open-weight models' raw capability.

7. Limitations and Future Research Directions

This study has material limitations. Sample sizes (~32–60 per condition) were constrained by API credit budget and time, and are too small to support confident claims about true population-level accuracy differences or statistical significance between conditions. The substitution of API-hosted Qwen/Gemma for genuinely local GGUF inference (necessitated by insufficient GPU resources) means this benchmark measures model-architecture performance under RAG, not the full sovereignty proposition of fully local, hardware-independent deployment — a follow-up study with actual on-premises inference is needed to close this gap. The anomalously low Gemini Pro scores were not diagnosed within the available time and likely reflect a pipeline artifact rather than a genuine capability finding. Finally, a deeper structural limitation is the scarcity of Africa-specific benchmark datasets, particularly in finance: unlike health (where AfriMed-QA exists), no mature Africa-grounded financial-LLM evaluation benchmark currently exists, forcing reliance on a single FinScope-derived Zindi competition

dataset; constructing a dedicated, larger-scale African financial-reasoning benchmark — analogous to AfriMed-QA's role in health — is itself a clear direction for future research.

References

African Development Bank. (2025). *Africa's AI productivity gain: Pathways to labour efficiency, economic growth and inclusive transformation*. African Development Bank Group.

Brookings Institution. (2025). *Foresight Africa 2025–2030: AI and emerging technologies*.

FinMark Trust. (2022). *FinScope MSME survey Zimbabwe 2022 findings*.

FinMark Trust. (2024). *FinScope MSME South Africa 2024: Key findings highlight urgent need for informal sector support*.

International Finance Corporation. (2025). *MSME finance*. World Bank Group.

Medical Device Network. (2026, January 21). *Davos26: Gates Foundation and OpenAI forge \$50m AI health partnership in Africa*.

Olatunji, T., Nimo, C., Owodunni, A., et al. (2024). *AfriMed-QA: A Pan-African, multi-specialty, medical question-answering benchmark dataset*. arXiv.

Reserve Bank of Zimbabwe. (2022). *Zimbabwe FinScope Consumer 2022 survey report*.

SME Finance Forum. (2025). *MSME finance gap*.