

Measuring AI Educational Equity Across Languages

A Four-Judge Evaluation of LLM Performance on South African Grade 12 Curriculum in English and Afrikaans

Aboobaker Cassim | University of the Witwatersrand, Johannesburg, South Africa
Global South AI Safety Hackathon 2026 — Africa Track (Evaluations & Benchmarks)

Abstract

Can a learner who asks an AI for help in their own language expect the same quality of help as one who asks in English? Evaluating Gemini 2.5 Flash on 129 South African Grade 12 examination questions posed in English and Afrikaans, this study finds a striking methodological result. The same mechanical keyword scorer, applied to the same fixed model responses, collapses the apparent equity gap by up to 478 times simply by switching the reference answer from English to Afrikaans. This is the central finding: the gap is an artefact of the reference language, not a property of the model. Two independent language-model judges, Gemini and Claude, corroborate it, finding negligible gaps of 0.006 to 0.038 across all subjects. As multilingual evaluation increasingly drives deployment decisions in the Global South, the scoring method can matter as much as the model measured.

1. Introduction

South Africa has eleven official languages, yet the artificial intelligence systems increasingly marketed as educational equalisers are built overwhelmingly on English text. AI tutoring tools are cheaper than human teachers, available at any hour, and reachable on an inexpensive mobile phone. For a country where access to quality teaching is unevenly distributed, this promise is genuine and important. But the promise rests on an unexamined assumption: that the tool works as well in the learner's language as it does in English.

Consider two Grade 12 learners preparing for the same national examination. One studies in English in a well-resourced Johannesburg school; the other studies in Afrikaans, or in Sepedi, in a rural town. Both turn to the same AI assistant the night before an exam. If the assistant answers the English learner accurately and the Afrikaans learner poorly, then the technology does not level the playing field — it tilts it further. The harm is not hypothetical or distant; it is delivered quietly, one homework question at a time, to the learners who can least afford it.

This study asks a direct question: when the same examination question is posed to a frontier language model in English and in Afrikaans, does the learner receive equivalent help? Afrikaans is deliberately chosen as a conservative test case. It is among the better-resourced of South Africa's indigenous-adjacent languages, with substantial digital text and a long history of formal instruction. If an equity gap appears even for Afrikaans, the gap for genuinely low-resource languages such as Zulu, Xhosa, or Sepedi would almost certainly be wider.

In the course of answering this question, a second and more consequential finding emerged — one not about the model at all, but about how we measure it. The apparent size of the equity gap depended almost entirely on the scoring method used, and in particular on the language of the reference answer against which responses were scored. This report argues that the dominant

approach to multilingual evaluation may be systematically misrepresenting how fairly AI systems treat speakers of other languages.

Our main contributions are:

- A reproducible bilingual evaluation pipeline that uses 129 official South African Grade 12 examination questions as ground truth, with the Department of Basic Education's verified marking memoranda as the reference standard.
- An empirical measurement of language-based equity gaps across three subjects chosen to span a range of language-dependency: Mathematics, Life Sciences, and Business Studies.
- A controlled demonstration that the apparent equity gap is an evaluation artefact: switching only the reference language of the keyword scorer, while holding all model responses fixed, collapses the gap by up to 478 times. This result is corroborated independently by two language-model judges.

2. Related Work

A growing body of work documents that large language models perform less well in African and other low-resource languages than in English. The Uhura benchmark of Bayes and colleagues evaluates models on scientific question answering and truthfulness across six typologically diverse African languages, including Northern Sotho (Sepedi), using human translation of existing English benchmarks, and reveals a significant performance gap between proprietary and open models. Broader efforts such as AfroBench, which spans many African languages and tasks, confirm that this gap is pervasive and correlates with the digital resource availability of each language. These results establish that a real capability gap exists at the level of factual accuracy.

This study departs from that literature in two ways. First, prior work largely relies on machine-translated or human-translated benchmark items adapted from English sources; here the evaluation data are official government examination papers, set bilingually and professionally by the Department of Basic Education, with both language versions legally equivalent and accompanied by official marking memoranda. This removes translation quality as a confounding variable and grounds the evaluation in the exact material a real learner would encounter.

Second, this work builds on the LLM-as-judge paradigm, in which a capable model evaluates the outputs of another model, an approach validated by Zheng and colleagues in their study of judging LLM-based assistants, who also catalogue its characteristic biases. Among these, the tendency of a model to favour its own outputs is documented directly by Panickssery and colleagues, and it is precisely this risk that motivates the inclusion of an independent third judge here. Where the LLM-as-judge literature is concerned chiefly with aligning automated judgments to human preferences, the present study uses multiple independent judges as a means of cross-examining the evaluation methodology itself. To our knowledge, no prior work isolates the reference language of a keyword scorer as the cause of an apparent language-equity gap, nor does any prior work use official curriculum marking memoranda as ground truth for a bilingual equity assessment. The relevance of evaluation-method bias to multilingual safety in the Global South

is argued by Banerjee and colleagues, whose agenda this study complements with concrete, reproducible measurement.

3. Methods

3.1 Data

The evaluation set comprises 129 questions drawn from three 2025 National Senior Certificate November examination papers: Mathematics Paper 1 (48 questions), Life Sciences Paper 1 (47 questions), and Business Studies Paper 1 (34 questions). These three subjects were chosen to span a spectrum of language-dependency. Mathematics is largely symbolic and computational; Life Sciences is heavily dependent on precise terminology and explanatory prose; Business Studies sits between, requiring applied reasoning in full sentences but with less specialised vocabulary than the sciences.

Each examination is published by the Department of Basic Education in both English and Afrikaans, with an accompanying marking memorandum in each language. For every question, the dataset records the English text, the Afrikaans text, the official correct answer, and the official marking steps, alongside metadata such as subject, mark allocation, and question type. The two language versions are professionally produced equivalents, so any measured difference in response quality reflects model behaviour rather than a difference in question difficulty. Questions that depended on diagrams or figures that could not be faithfully represented in text were excluded from the Life Sciences set to avoid confounding the language comparison with visual comprehension.

3.2 Generation

Each question was submitted to Gemini 2.5 Flash via the google-genai SDK, once in English and once in Afrikaans, yielding 258 generated responses. A fixed system prompt framed the model as a tutor assisting a South African Grade 12 learner and instructed it to answer clearly, show working where appropriate, and respond in the same language as the question. Generation used a temperature of zero for determinism, with a one-second delay between calls. The instruction to remain in the language of the question was respected in practice: Afrikaans questions produced genuinely Afrikaans responses using correct subject terminology, rather than translated English.

3.3 Four Judges

Each of the 258 responses was scored by four methods, chosen so that their disagreements would be diagnostic. The first two share an identical mechanical scorer and differ in one variable alone, which makes their comparison a controlled experiment.

Judge 1 — Keyword overlap against the English memorandum. A purely mechanical scorer that measures the fraction of reference tokens from the marking memorandum that appear in the response, after lowercasing and stripping punctuation. The combined score weights the memo answer at sixty percent and the marking steps at forty percent. Here the reference memorandum is in English, so Afrikaans responses are compared against English tokens. This mirrors the standard practice of scoring multilingual output against an English key, and is included precisely so its behaviour can be examined.

Judge 1b — Keyword overlap against the Afrikaans memorandum. This is the controlled experiment at the heart of the study. The model responses, the scoring formula, and the tokenisation are all held completely fixed between Judge 1 and Judge 1b; the only thing that changes is the language of the reference answer, which is now the official Afrikaans memorandum rather than the English one. Because every other factor is identical — the Afrikaans responses scored are byte-for-byte the same in both conditions — any difference in the measured equity gap can have no other explanation than the reference language. This is what makes the finding conclusive rather than merely suggestive.

Judge 2 — Gemini as self-judge. Gemini 2.5 Flash is shown the question, the official answer, the marking steps, and the response, and asked to return a single quality score between zero and one. Because the model judges its own output, it may be generous to itself — a tendency Panickssery and colleagues document as LLM evaluators recognising and favouring their own generations. This is a known limitation, mitigated by the independent fourth judge.

Judge 3 — Claude as independent judge. Claude Haiku is given the same scoring prompt and returns a score between zero and one. As a model from a different developer with no stake in Gemini's output, it provides an independent check on both the keyword scorer and Gemini's self-assessment.

In total the study comprises approximately 258 generation calls and roughly 774 judging calls, just over one thousand model interactions. All code, data, and raw outputs are available in the public repository so that the full pipeline can be reproduced.

4. Results

Table 1 reports mean scores for English and Afrikaans and the equity gap (English minus Afrikaans) for every combination of subject and judge. A positive gap indicates that English was scored higher.

Subject	Judge	English	Afrikaans	Gap
Mathematics	Judge 1 — Keyword (English memo)	0.671	0.469	0.202
	Judge 1b — Keyword (Afrikaans memo)	0.671	0.675	0.004
	Judge 2 — Gemini (self-judge)	0.980	0.972	0.008
	Judge 3 — Claude (independent)	0.850	0.845	0.006
Life Sciences	Judge 1 — Keyword (English memo)	0.736	0.208	0.529
	Judge 1b — Keyword (Afrikaans memo)	0.736	0.728	0.008
	Judge 2 — Gemini (self-judge)	1.000	0.972	0.028

Subject	Judge	English	Afrikaans	Gap
	Judge 3 — Claude (independent)	0.941	0.903	0.038
Business Studies	Judge 1 — Keyword (English memo)	0.553	0.219	0.334
	Judge 1b — Keyword (Afrikaans memo)	0.553	0.554	0.0001
	Judge 2 — Gemini (self-judge)	0.841	0.832	0.008
	Judge 3 — Claude (independent)	0.747	0.764	-0.017

Table 1. Mean scores and equity gap by subject and judge. *Judge 1 gaps are large because Afrikaans responses are compared against English reference text. Judge 1b uses the official Afrikaans memorandum and the gap collapses in every subject. Both LLM judges independently confirm the same result.*

The decisive comparison is between Judge 1 and Judge 1b. Holding the model responses fixed and changing nothing but the reference language, Judge 1b reduced the apparent equity gap relative to Judge 1 by approximately 50 times in Mathematics (from 0.202 to 0.004), approximately 66 times in Life Sciences (from 0.529 to 0.008), and approximately 478 times in Business Studies (from 0.334 to 0.0001). These reductions are achieved with no change whatsoever to the model responses — the Afrikaans answers are byte-for-byte identical in both conditions. The apparent inequity therefore lived entirely in the language of the reference key, not in the model's behaviour.

The two language-model judges corroborate this independently. Under Judge 1, Afrikaans appears markedly worse than English in every subject, with gaps ranging from 0.20 in Mathematics to 0.53 in Life Sciences. Yet when the identical responses are assessed by either Gemini or Claude, those gaps very nearly vanish, falling between 0.006 and 0.038. Two judges, one of them from a different developer entirely, agree with the Afrikaans-memo keyword scorer that the responses are of comparable quality across the two languages.

Two further observations reinforce the interpretation that the Judge 1 gap is an artefact. First, the Afrikaans keyword scores under Judge 1 are compressed into a narrow band (0.21 to 0.47) regardless of subject, whereas the English keyword scores rise with subject (0.55 to 0.74). This is the signature of a ceiling imposed by the scorer rather than by the model: Afrikaans responses cannot accumulate credit because they are matched against English tokens, no matter how correct they are. Second, the residual gap that does survive independent judging is largest in Life Sciences (0.028 to 0.038), the most terminology-dependent subject, which is exactly where a small genuine difference in specialised Afrikaans vocabulary would be expected to appear. The independent judges thus detect a faint, plausible signal precisely where one should exist, while the English-memo keyword scorer reports large gaps everywhere.

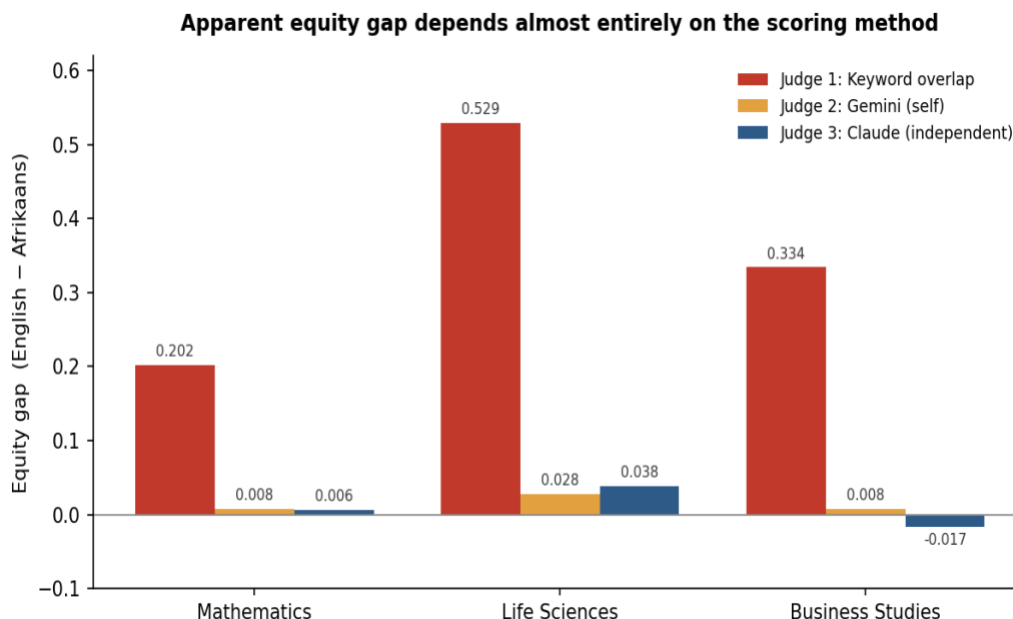


Figure 1. Equity gap (English – Afrikaans) by subject under each judge. The English-memo keyword scorer reports large gaps in every subject, while both LLM judges report gaps near zero — and in one case a slight Afrikaans advantage. The disagreement between methods is far larger than any disagreement between languages.

5. Discussion and Limitations

The Judge 1 to Judge 1b comparison is a controlled experiment that isolates the reference language as the sole cause of the apparent equity gap. Because the model responses, the scoring formula, and the tokenisation are all held fixed, and only the language of the reference answer changes, the collapse of the gap — by up to 478 times — cannot be attributed to anything else. The apparent inequity was not a property of how Gemini treats Afrikaans learners; it was a property of measuring Afrikaans answers against an English key. Keyword-overlap scoring rewards surface vocabulary matching, so a correct Afrikaans answer that uses correct Afrikaans terminology is penalised simply for not containing English words. When the same scorer is instead pointed at the official Afrikaans memorandum, or when evaluation asks whether the response actually answers the question — the judgment both LLM judges make — the apparent inequity disappears.

This matters for AI safety in a concrete way. Evaluation results drive deployment decisions, procurement, and policy. If the instruments used to certify that a system performs adequately in a given language are themselves biased against that language, then two distinct failure modes follow. A system may be wrongly judged unfit for a language in which it is in fact competent, denying communities a useful tool; or, conversely, a flawed evaluation may be trusted in one direction and not examined in the other, so that genuine harms go undetected because the measurement instrument is not trusted to begin with. In either case, the community that speaks the non-English language bears the cost of a measurement error it had no part in making. For the Global South, where multilingual deployment is the norm rather than the exception, the reliability of the evaluation method is a first-order safety concern.

It would be a mistake to read this result as evidence that language inequity in AI does not exist. The established literature documents real capability gaps, and the small residual gap this study finds in Life Sciences is consistent with that. The claim here is narrower and more specific: a widely used class of evaluation method exaggerates those gaps to the point of distortion, and conclusions drawn from such methods should be treated with caution.

5.1 Limitations

Several constraints bound these findings. The study evaluates a single model, Gemini 2.5 Flash, and a single examination year, 2025; the results may not generalise to other models or cohorts. The English-only keyword scorer is both the bias under examination and a faithful reproduction of common practice, so its results should be read as a demonstration of that practice's failure mode rather than as an independent measurement. Afrikaans is a relatively well-resourced language, so the near-zero gaps found here cannot be assumed to hold for Zulu, Xhosa, or Sepedi, where both genuine capability gaps and evaluation artefacts may be larger. The two LLM judges are not infallible: Judge 2 grades Gemini's own output and may be self-favourable, and although Judge 3 is independent, LLM judges as a class can carry their own systematic preferences. Diagram-dependent questions were excluded from Life Sciences, which may narrow the representativeness of that subject. Finally, no native Afrikaans speaker validated the responses; human evaluation remains the gold standard and its absence is the most important gap in the present design.

5.2 Dual-Use Considerations

The pipeline could be misused. Because it demonstrates that LLM-judge scores are far more favourable than English-memo keyword scores, a developer could selectively adopt LLM judging to present a system as more equitable than a more demanding evaluation would suggest, masking genuine gaps. Conversely, the finding could be over-generalised into a claim that AI is already equitable and that no further investment in low-resource languages is needed, which the data do not support. The keyword scorer's partial validity — it does capture something real about vocabulary coverage when the reference is in the right language — should not be dismissed wholesale. The responsible reading is that no single automated method suffices and that method triangulation, ideally anchored by human judgment, is necessary.

5.3 Future Work

The natural next steps are to extend the evaluation to genuinely low-resource South African languages such as Zulu, Xhosa, and Sepedi; to add further frontier models including GPT-4o and Llama so that the methodological finding can be tested for generality; to incorporate multiple examination years for temporal robustness; and to introduce native-speaker validation as a fifth, human judge. Integrating the pipeline with an established evaluation framework such as the UK AI Safety Institute's Inspect would make the method portable, and a hybrid scorer that combines keyword coverage with LLM judgment could capture the legitimate signal in each while discarding the artefacts.

6. Conclusion

This study set out to ask whether a frontier language model helps an Afrikaans-speaking learner as well as an English-speaking one. The answer, established by a controlled experiment and corroborated by two independent judges across 129 official examination questions, is that Gemini 2.5 Flash performs equitably across the two languages. The keyword-overlap scoring that dominates much multilingual evaluation reports gaps up to 478 times larger than the genuine signal — for no reason other than that its reference answers are written in English. Point the very same scorer at the official Afrikaans memorandum, changing nothing about the model's responses, and the gap collapses to near zero.

The deeper lesson is that, in multilingual AI evaluation, the choice of measurement instrument can matter as much as the system being measured. An evaluation method that compares non-English answers against English keys does not measure inequity so much as manufacture it. As AI systems are deployed across the many languages of the Global South, getting the measurement right is not a technical afterthought but a precondition for fair and safe deployment. The learners who stand to gain most from these tools are precisely those whose languages are most likely to be mismeasured, and they deserve evaluation methods worthy of the trust placed in them.

Code and Data

Code repository: github.com/AbCassim04/llm-equity-eval

Data source: South African Department of Basic Education, 2025 November NSC Examination Papers, education.gov.za. The repository contains the extracted bilingual question set, all raw model responses, and the scored outputs for every judge, enabling full reproduction of the results.

LLM Usage Statement

Claude was used to assist with project scoping, evaluation-pipeline architecture, debugging, and drafting of this report. All experimental results were generated by the pipeline described above, independently verified by the author, and are fully reproducible from the public repository. The research design, the analysis, and the conclusions are the author's own.

References

- Bayes, E., Azime, I. A., Alabi, J. O., Kgomo, J., Eloundou, T., Proehl, E., Chen, K., Khadir, I., Etori, N. A., Muhammad, S. H., Mpanza, C., Thete, I. P., Klakow, D., & Adelani, D. I. (2024). Uhura: A Benchmark for Evaluating Scientific Question Answering and Truthfulness in Low-Resource African Languages. *arXiv:2412.00948*.
- Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., & Stoica, I. (2023). Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *Advances in Neural Information Processing Systems 36 (NeurIPS 2023), Datasets and Benchmarks Track*. *arXiv:2306.05685*.
- Panickssery, A., Bowman, S. R., & Feng, S. (2024). LLM Evaluators Recognize and Favor Their Own Generations. *arXiv:2404.13076*.

- Banerjee, S., Hazra, R., & Mukherjee, A. (2026). Bridging the Multilingual Safety Divide: Efficient, Culturally-Aware Alignment for Global South Languages. arXiv:2602.13867.
- Adelani, D. I., et al. (2025). AfroBench: How Good are Large Language Models on African Languages? Findings of the Association for Computational Linguistics (ACL 2025). arXiv:2311.07978.
- South African Department of Basic Education (2025). National Senior Certificate November Examination Papers and Marking Guidelines. Available at education.gov.za.
- UK AI Safety Institute (2024). Inspect: An Open-Source Framework for Large Language Model Evaluations. Available at inspect.aisi.org.uk.