Interpreting isiZulu Idioms with Large Language Models: A Comparative Study of Zero-Shot Prompting, Retrieval-Augmented Generation, and Supervised Fine-Tuning

Nontokozo Buhle Ntandoyenkosi Manukuza[1], **Vukosi Marivate**[1,2,3], **Tshephisho Joseph Sefara**[4], **Sthembiso Mkhwanazi**[1], **Lungile Mncwango**[5], **Kwanele Radele**[1]

[1]Data Science for Social Impact, Department of Computer Science, University of Pretoria, South Africa
[2]Lelapa AI, South Africa
[3]African Institute of Data Science and Artificial Intelligence, University of Pretoria, South Africa
[4]Council for Scientific and Industrial Research, Pretoria, South Africa
[5]Department of African Languages, University of Pretoria, South Africa
u22832191@tuks.co.za, vukosi.marivate@cs.up.ac.za, tsefara@csir.co.za

Abstract

Idiom interpretation requires cultural knowledge, context, and pragmatic understanding that cannot be inferred from literal word meanings alone. This challenge is particularly important for isiZulu, a richly inflected Nguni Bantu language with limited representation in the web-scale corpora used to train modern large language models (LLMs). We compare three approaches to isiZulu idiom interpretation (zero-shot prompting, retrieval-augmented generation (RAG) with glossary evidence, and supervised fine-tuning on a purpose-built, context-rich corpus) across five open-weight LLMs, evaluating them on 323 held-out idioms. Performance is evaluated using automated metrics and human assessment by expert isiZulu speakers. Supervised fine-tuning provides the strongest overall improvement, increasing the mean human evaluation score from 2.61 to 3.68 out of 5. RAG also improves interpretation quality, although its gains are smaller and are not fully captured by conventional surface-level metrics. Across the fifteen model–condition systems, COMET aligns best with human assessment, whereas embedding cosine similarity is the only metric that does not align positively with it, suggesting that this metric should be validated against human judgement before being used to evaluate idiomatic interpretation in low-resource languages. We contribute the isiZulu Idiom Dataset, a morphology-aware retrieval pipeline, and a reusable protocol for evaluating LLM-based idiom interpretation in low-resource African languages.

1 Introduction

When a fluent isiZulu speaker says that someone *ucele empunzini* (literally, "he or she asked from the duiker"), they are not talking about the small antelope at all. They mean the person fled at great speed, in fear or in urgent flight from danger: the image draws on the duiker's habit of darting away the instant it is threatened, as though the person had borrowed the animal's speed to escape. No word-by-word translation recovers this. A reader who knows only the dictionary senses of *ukucela* ("to ask") and *impunzi* ("duiker") misses the meaning entirely, because it resides not in the words but in the shared cultural knowledge of the community that uses them. Asked to interpret such an expression, a large language model (LLM) confronts exactly the gap between surface form and conveyed meaning that idioms exploit.

The gap is acute for isiZulu, a Nguni Bantu language spoken at home by about fifteen million South Africans yet supported by only a thin computational footprint (Joshi et al., 2020; Adebara and Abdul-Mageed, 2022). Idioms are frequent in speech, culturally dense, and almost absent from web-scale corpora, so they are the domain in which the low-resource gap bites hardest. Despite growing momentum in African-language natural language processing (NLP) (Nekoto et al., 2020), figurative language remains largely unaddressed for Bantu languages. Two language properties compound the difficulty: *agglutinative morphology* packs concords, tense, aspect, and derivational affixes into single forms, so an idiom's decisive element may be a verbal root buried under several affixal layers (Marquard et al., 2025; Tsotetsi et al., 2025); and *conjunctive orthography* writes these complexes as unbroken words, defeating tokenisation that splits text at spaces. The problem is not *only* data volume, however: even with a large corpus, idiom interpretation would remain hard, because the way figurative meaning is stored and

This work is licensed under CC BY SA 4.0. To view a copy of this license, visit https://creativecommons.org/licenses/by-sa/4.0/deed.en
The copyright remains with the authors.

activated resists the compositional assumptions of current architectures (Section 2.1). This separates failure for lack of *exposure* from failure for lack of the right *representation*.

We therefore compare three complementary strategies, each targeting a different part of the problem. *Zero-shot prompting*[1] asks the model cold, with no examples or extra information, and so tests whatever figurative competence it already has. *Retrieval-augmented generation* (RAG)[2] is akin to handing the model an open dictionary: relevant glossary entries are looked up and placed in the prompt before it answers. *Supervised fine-tuning*[3] gives the model additional training lessons on idiom–interpretation pairs, changing its internal parameters. To our knowledge, no systematic comparison of these strategies exists for isiZulu idiom interpretation, nor any analysis of how a model's pre-training interacts with each in a morphologically complex, low-resource context. We close that gap with a controlled, fully crossed evaluation of five open-weight LLMs across all three conditions on a shared held-out test set, scored by automated metrics and expert human judgement. This paper makes four contributions:

- **A resource.** The *isiZulu Idiom Dataset*: an expert-annotated dataset of 3,233 entries, each with a figurative interpretation and usage examples; no sub-Saharan Bantu idiom resource appears in prior figurative-language benchmarks.

- **A pipeline.** A morphologically informed retrieval pipeline (a bilingual glossary indexed with FAISS (Facebook AI Similarity Search), a morphological decomposition module, and an LLM-based word-sense disambiguation (WSD) layer), together with QLoRA (quantised low-rank adaptation) fine-tuning protocols, to be released where permitted.

- **An empirical comparison.** The first evaluation of these strategies for a low-resource African language, anchored by reliable expert human judgement and examining how pre-training history shapes the return on each.

- **A metric-reliability finding.** Evidence from the fifteen model–condition systems that COMET tracks expert judgement best, while embedding cosine similarity is the only metric whose rank correlation with it is negative ($\tau = -0.251$; not significant), a direction supported by an item-level test.

Because this work spans NLP and the humanities, every abbreviation is spelled out at first use, and Appendix A collects them in a glossary with links to introductory explanations. Section 2 situates the work in the psycholinguistics of idioms and prior low-resource NLP. Section 3 describes the dataset, the three conditions, and the evaluation protocol. Section 4 reports the automated, retrieval, and human results and tests which automated metrics can stand in for human judgement. Section 5 interprets the findings and draws out their implications for evaluation, resourcing, and cultural-heritage work.

2 Related Work

2.1 Why idioms resist language models

Psycholinguistic research both explains why idiom interpretation is hard and motivates our three strategies. The *lexical representation hypothesis* (Cacciari and Tabossi, 1988; Swinney and Cutler, 1979) holds that idioms are stored as multi-word chunks, and the *direct access model* (Gibbs, 1980) shows that idioms in supportive discourse are processed no slower than literal paraphrases, so context is integral to activating figurative meaning; this is the warrant for RAG. The *configuration hypothesis* (Titone and Connine, 1999) adds that comprehension proceeds compositionally until a salient *key word* triggers the idiomatic reading, with word meanings still contributing even for conventional idioms (Cacciari and Glucksberg, 1991). In isiZulu that key word may be a verbal root buried under agglutinative morphology, so a tokeniser that cannot recover roots misses the trigger *even if the idiom was seen in training*, the warrant for our morphological-decomposition module. Finally, idiomaticity is scalar (Nunberg et al., 1994): some idioms are partly transparent, others fully opaque. The dataset spans this range, which our multi-dimensional human rubric respects.

2.2 Idiomatic competence in LLMs

Two failure modes recur in the LLM idiom literature. The first is literal interpretation when context

[1] https://en.wikipedia.org/wiki/Zero-shot_learning

[2] https://en.wikipedia.org/wiki/Retrieval-augmented_generation

[3] https://en.wikipedia.org/wiki/Fine-tuning_(deep_learning)

is absent (Mi et al., 2025), a *grounding* deficit that our RAG condition targets. The second is that competence on familiar idioms often reflects memorisation rather than interpretation (Kim et al., 2025), a *representation* deficit that our fine-tuning condition targets. Both intensify as resources diminish (Almheiri et al., 2026; Sakhawat et al., 2026). IsiZulu, whose idiomatic data are almost absent from pre-training corpora, is thus a demanding test case: with the memorisation heuristic unavailable, observed competence reflects genuine interpretation, which the zero-shot condition quantifies before either intervention.

Two structural gaps motivate this study. First, annotated idiom resources remain concentrated in high-resource languages. Recent multilingual benchmarks widen coverage but still exclude sub-Saharan Bantu: MIDAS (Kim et al., 2025) pairs idioms with meanings across six languages including agglutinative Turkish, yet contains no African language, and the most extensive dataset survey to date (Flor et al., 2025) identifies no annotation framework, benchmark, or pipeline for any African language. Second, figurative-generation evaluation typically relies on surface-overlap or embedding metrics whose validity for this task is assumed rather than established; agreement between automatic metrics and human judgement is seldom quantified in a low-resource, culturally grounded setting. We address the first gap with a purpose-built dataset and the second with an explicit metric-reliability analysis.

2.3 Parameter-efficient fine-tuning

Low-Rank Adaptation (LoRA) (Hu et al., 2022) and its quantised variant QLoRA (Dettmers et al., 2023) make task-specific adaptation of billion-parameter models feasible on modest hardware. Instead of updating every weight, they train a small set of add-on "adapter" matrices on top of a frozen backbone, which QLoRA additionally compresses to 4-bit precision to save memory. These methods suit low-resource settings, where full fine-tuning is impractical and small curated datasets make over-fitting a real risk. QLoRA in particular allows all five of our models, including a 20-billion-parameter model, to be fine-tuned on a single graphics processing unit (GPU).

2.4 Retrieval-augmented generation

RAG (Lewis et al., 2020) augments the knowledge stored in a model's parameters with evidence retrieved at inference time. Its effectiveness in low-resource, morphologically rich languages turns on two determinants. On *architecture*, morphology-aware retrievers built for agglutinative Bantu languages substantially outperform generic multilingual baselines (Nzeyimana and Rubungo, 2025), so retrieval must accommodate noun-class morphology rather than rely on generic subword matching. On *corpus quality*, unfiltered corpora degrade generation and non-native embeddings inject retrieval noise (Wu et al., 2024), so we pair the retriever with a curated, expert-verified glossary rather than scraped text; culturally grounded prompting over retrieved evidence has likewise proven effective for figurative tasks in other low-resource languages (Thakur, 2025). Whether high-quality retrieval yields downstream gains ultimately depends on the model's ability to *integrate* evidence, and our intrinsic retrieval evaluation helps to separate this ability from coverage.

3 Methodology

We build and evaluate three conditions (a zero-shot protocol, a RAG pipeline, and a set of fine-tuned variants) for five models on a single held-out test set, so that every condition is scored on identical items. The study is guided by four research questions (RQs):

RQ1 How accurately, fluently, and culturally appropriately can current LLMs interpret isiZulu idioms zero-shot, by automatic and human evaluation?

RQ2 Does RAG over a morphologically informed bilingual glossary improve interpretation quality relative to the zero-shot baseline?

RQ3 Does QLoRA fine-tuning improve quality relative to baseline, and how does it compare with RAG?

RQ4 How do base-model characteristics influence the effectiveness of each strategy?

3.1 Models

The five models are chosen to span deliberately different design choices (African-language pre-training, parameter scale, instruction tuning and architectural lineage): **AfroLlama V1** (8B, i.e. 8 billion parameters) is continually pre-trained on African-language data; **Meta-Llama 3.1 8B** is a strong general-purpose open-weight base model;

Meta-Llama 3.1 8B Instruct is its instruction-aligned variant; **GPT-OSS 20B** is a substantially larger open-weight model; and **Gemma 3 12B** represents a distinct architectural lineage. Because these properties vary jointly across the five models (even the Llama 3.1 base/Instruct pair differs in both instruction tuning and baseline inference precision, FP16 (16-bit floating point) versus 4-bit), per-model differences are reported as observed associations, not as causal effects of any one property.

3.2 The isiZulu Idiom Dataset

The dataset comprises 3,233 idiomatic expressions drawn from authoritative lexicographic sources, informed by linguistic studies of isiZulu idiom (Khumalo and Dlamini, 2025), and verified by a panel of four adult native isiZulu speakers external to the research team (sources in Appendix B). Each entry records the idiom surface form, its gold figurative interpretation, and at least one usage example fixing the pragmatic context. Because idiomaticity is scalar, the corpus spans partly transparent through fully opaque idioms, letting the evaluation distinguish models that memorise glosses from those that recover compositional cues.

The 3,233 idioms are shuffled under a fixed seed (42) and split 80/10/10, identically for all five models, into a *training set* of 2,587 idioms, the only items used for parameter updates; a *validation set* of 323, used for the hyperparameter search and best-checkpoint selection; and a held-out *test set* of 323, on which every model and condition is evaluated and every reported result is computed. The three sets do not overlap, and the test set is never seen during training, tuning, or checkpoint selection, so condition-level differences reflect adaptation rather than item difficulty.

3.3 Condition 1: Zero-shot baseline

Each model interprets each test idiom using only its parametric knowledge, under a fixed isiZulu prompt requesting a figurative interpretation and usage examples: no external evidence and no task-specific tuning, so any correct interpretation must come from competence latent in pre-training (RQ1). This sets the reference point against which both adaptation strategies are measured. The same zero-shot prompt, with no worked examples and no model-specific tailoring, is applied identically to all five models (Appendix C).

3.4 Condition 2: Retrieval-augmented generation

The RAG pipeline proceeds in four stages (Figure 1). (i) *Morphological decomposition* segments each idiom into candidate roots and affixes, exposing the verbal roots that conjunctive orthography and subword tokenisation obscure. (ii) *Retrieval* searches a bilingual isiZulu–English glossary indexed with FAISS (Johnson et al., 2019), a library for fast nearest-neighbour search over vector embeddings.[4] Glossary entries are embedded with a multilingual encoder so that related forms lie near one another, and the top-k ($k = 3$) candidate glosses per component are returned by cosine similarity. (iii) A *word-sense disambiguation* (WSD) layer (WSD being the task of choosing which of a word's several meanings is intended in context)[5] resolves sense among the candidates. This is essential because agglutinative roots are frequently polysemous; for each word the layer either selects one candidate (SELECT) or rejects them all and supplies the meaning itself (REJECT), while words with only one retrieved candidate are marked SINGLE upstream and bypass the model. Its success rate is reported as the WSD hit rate. (iv) The resolved glosses are supplied to the generating model with a *chain-of-thought* prompt and three worked exemplars, which asks it to reason step by step from literal word meanings to the literal image to the figurative meaning.

The decomposition module is not a full finite-state analyser but a lightweight, lexically grounded segmenter: candidate roots come from rule-based prefix and suffix stripping, FastText subword matching against the glossary vocabulary, and fuzzy (edit-distance) matching for orthographic and inflectional variants. It deliberately favours candidate coverage over a single hard segmentation, since the WSD layer makes the final sense decision. The retriever uses `intfloat/multilingual-e5-large` (1024-dimensional embeddings), which covers the closely related isiXhosa but reaches isiZulu only through cross-lingual transfer. The WSD layer is a separate, smaller Gemma model (`gemma:2b`, not one of the five evaluated models) at temperature 0.3, held fixed across all runs, so the evidence supplied to each generator is identical and cross-

[4] https://github.com/facebookresearch/faiss

[5] https://en.wikipedia.org/wiki/Word-sense_disambiguation

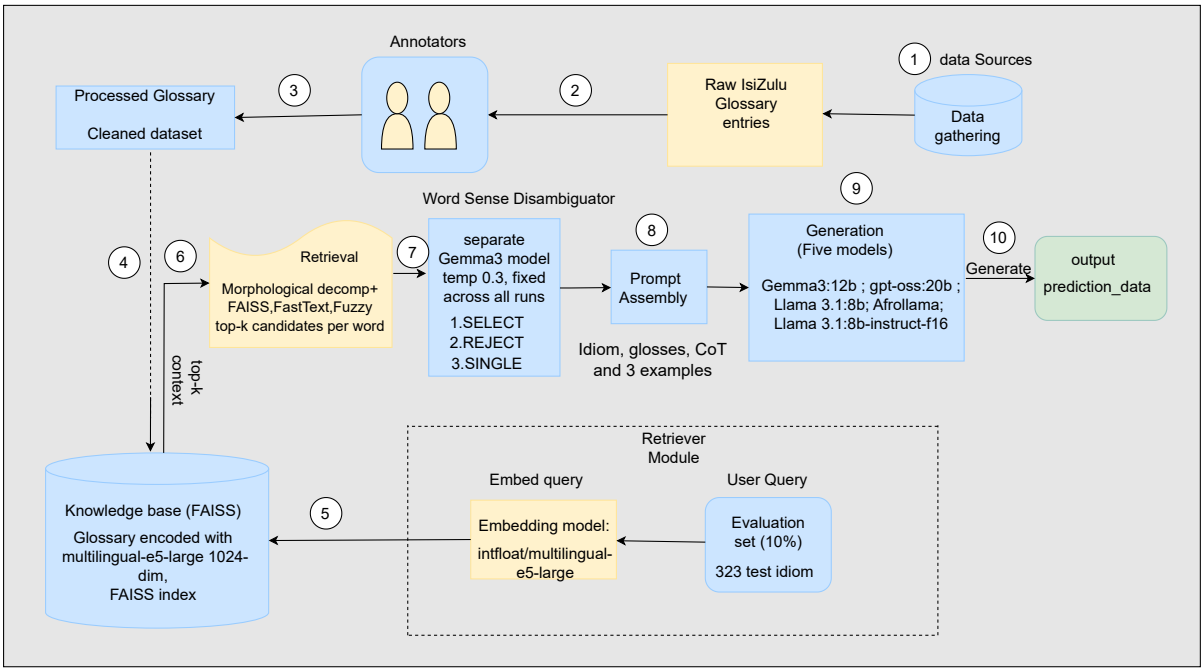

Figure 1: RAG pipeline for isiZulu idiom interpretation.

model differences reflect *integration* rather than disambiguation. Appendices F and G give the decomposition rules and the verbatim prompts; code will be released where permitted.

3.5 Condition 3: Supervised fine-tuning

Each model is fine-tuned on the 2,587-idiom training set (Section 3.2; Figure 2) using QLoRA: LoRA adapters ($r = 16$) are trained over a backbone quantised to 4-bit NormalFloat (NF4), with the same adapter configuration for all five models (apart from dropout; Appendix E). Hyperparameter optimisation (HPO) uses Optuna's tree-structured Parzen estimator (TPE), a Bayesian search method, to minimise loss on the 323-idiom validation set over a fixed budget of ten trials per model (five epochs each) and an identical search space (Appendix E); this uniformity keeps model differences from being confounded with differences in the searched region. Baseline and fine-tuned generation is greedy (deterministic), except GPT-OSS 20B at baseline, which uses temperature 0.7 because greedy decoding produced degenerate empty responses; RAG generation samples at temperature 0.7 for all five models, with a larger token budget. The baseline-to-fine-tuning comparison is therefore deterministic for four of five models (see Limitations).

Because the *Isisho/Incazelo/Izibonelo* (idiom / explanation / examples) output scaffold is learned during training, the fine-tuned models are prompted with only the bare instruction and the target idiom. The fine-tuned condition therefore receives strictly *less* prompt context than the baseline, so its gains cannot be an artefact of a richer prompt.

3.6 Evaluation protocol

Automated metrics. Five metrics in three tiers triangulate different notions of quality. *Surface overlap*: chrF++ (Popović, 2017), a character-and-word n-gram F-score, well suited to agglutinative morphology where word-level metrics such as BLEU (Bilingual Evaluation Understudy) fragment. *Embedding similarity*: BERTScore F1 (Zhang et al., 2020) (computed with XLM-R-large), which matches contextual token embeddings of prediction and reference, and sentence-level embedding cosine similarity. *Learned metrics trained on human judgements*: COMET (Crosslingual Optimized Metric for Evaluation of Translation) (Rei et al., 2020), trained on human direct-assessment ratings of translation, and AfriCOMET (Wang et al., 2024), which retrains COMET on human ratings for African languages using an African-centric encoder (Alabi et al., 2022). All are applied off-label to a monolingual, non-translation task and are therefore validated against human judgement rather than trusted as ground truth. SSA-COMET (Sub-Saharan Africa COMET) (Li et al., 2025), released after our metric

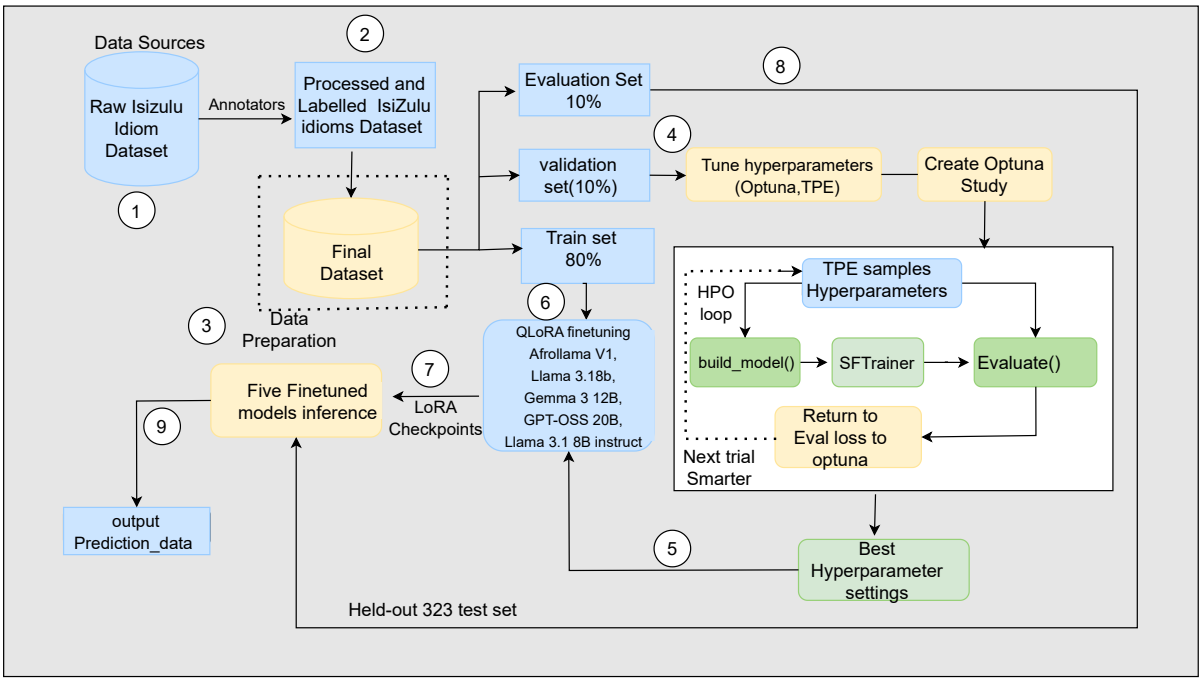

Figure 2: Fine-tuning pipeline for isiZulu idiom interpretation.

outputs were finalised, is the natural next metric to validate.

Expert human evaluation. A panel of three expert isiZulu speakers rated outputs on a Direct Assessment (DA) rubric (absolute ratings of each output rather than rankings) with five independent 1–5 dimensions: Fluency, Adequacy, Cultural Fidelity, Completeness, and Coherence (Appendix H). A multi-dimensional rubric is needed because idiom interpretation fails in qualitatively different ways: fluent but wrong, correct but culturally flat, adequate but incomplete. After removing 123 duplicate submissions (keeping each rater's latest rating per item; no rating was altered or reconciled toward consensus), the set comprises 1,800 ratings and 1,473 free-text comments. The same 40 idioms, drawn from the 323-item test set, were used for every model and every condition, so each of the 15 model×condition cells contains 40 outputs rated by all three experts (120 ratings per cell), and all human ratings are directly comparable on an identical item set. Ratings were not blind to model or condition, and the raters are co-authors (see Limitations). The human ratings serve both as ground truth on interpretation quality and as the reference for testing each automated metric.

4 Results

4.1 Automated results

Addressing RQ2 and RQ3, supervised fine-tuning gives the strongest automated performance overall (Table 1): it raises mean COMET, AfriCOMET, BERTScore, and chrF++, whereas RAG produces smaller, metric-dependent changes. Cosine similarity behaves differently, rising under RAG but falling after fine-tuning; we return to this divergence in Section 4.4.

Disaggregating by model exposes the base-model pattern that RQ4 addresses: the African-pretrained model has the strongest automated baseline, while gains from adaptation are largest for the models that start weakest. AfroLlama V1 holds the highest baseline AfriCOMET (0.411), far above the 0.256 and 0.249 of the two vanilla Llama models, a pattern consistent with African-language pre-training conferring a head start, although pre-training is not varied independently of other model properties in this design. The vanilla Llama models gain most from fine-tuning: Llama 3.1 8B Instruct lifts COMET from 0.563 to 0.726 and AfriCOMET from 0.249 to 0.361, closing almost the entire COMET gap to AfroLlama (0.726 versus 0.731), although a gap remains on AfriCOMET (0.361 versus 0.411). chrF++, the metric most sensitive to isiZulu morphology, is an exception: its largest fine-tuning gain is for AfroLlama V1 ($14.40 \rightarrow 19.26$),

Table 1: Per-model automated results across five metrics and three conditions (B = baseline, R = RAG, F = fine-tuned); higher is better. Best per model/metric in **bold**.

	COMET			AfriCOMET			BERTScore F1			chrF++			Cosine sim.		
Model	B	R	F	B	R	F	B	R	F	B	R	F	B	R	F
AfroLlama V1	0.703	0.714	**0.731**	0.411	**0.415**	0.391	0.840	0.847	**0.855**	14.40	14.49	**19.26**	0.698	**0.726**	0.640
Llama 3.1 8B	0.575	0.599	**0.727**	0.256	0.252	**0.359**	0.835	0.841	**0.850**	13.48	13.87	**16.93**	0.668	**0.708**	0.640
Llama 3.1 8B-It	0.563	0.594	**0.726**	0.249	0.255	**0.361**	0.834	0.839	**0.848**	13.96	14.15	**16.11**	0.689	**0.707**	0.640
GPT-OSS 20B	0.619	0.601	**0.658**	**0.369**	0.356	0.351	0.817	0.802	**0.838**	10.98	8.23	**13.32**	0.722	**0.733**	0.673
Gemma 3 12B	0.620	0.662	**0.678**	**0.360**	0.340	0.294	0.828	**0.844**	0.835	12.96	**14.30**	12.59	**0.729**	0.709	0.600
Mean	0.616	0.634	**0.704**	0.329	0.324	**0.351**	0.831	0.835	**0.845**	13.16	13.01	**15.64**	0.701	**0.717**	0.639

the strongest baseline model. Two models break the overall pattern. GPT-OSS 20B is *degraded* by retrieval, recording the study's lowest chrF++ (8.23) and BERTScore F1 (0.802). Gemma 3 12B shows the largest AfriCOMET drop under fine-tuning ($0.360 \rightarrow 0.294$) and is the only model whose RAG scores beat its fine-tuned scores on BERTScore F1 and chrF++, an anomaly that reappears in the human data.

4.2 Retrieval quality

On a hand-verified gold sample of 100 component words, the retriever placed the correct lemma first for 71% of words (recall@1 $= 0.71$) and within the top three for 86% (recall@3 $= 0.86$), with a mean reciprocal rank (MRR) of 0.773. Across the full test set, the WSD stage resolved a near-identical proportion of words for every model (hit rate 70.7–71.9%). Because coverage is high and effectively model-independent, every generator receives essentially the same evidence, and variation in RAG's effect across models is more plausibly attributable to how each model *integrates* that evidence than to retrieval coverage.

4.3 Human evaluation

Inter-rater agreement. Reliability is measured with Krippendorff's α (Krippendorff, 2004), a chance-corrected agreement coefficient where 1 means perfect agreement and 0 chance-level agreement;[6] we use the interval distance, which penalises large disagreements more (Appendix I). Over the 600 fully crossed outputs the panel is highly reliable: $\alpha = 0.890$ on the overall score, with all five per-dimension values in $[0.829, 0.882]$ (ordinal α: 0.822–0.868), above the 0.80 threshold conventionally taken as strong agreement. Per-rater means lie within a narrow 0.21-point band (3.12–3.33), so no rater is dominantly lenient or severe. The five dimensions are related but separable (pairwise Spearman ρ from 0.75 to 0.86; Appendix J).

Findings. The overall rating rises from 2.61 at baseline to 3.30 under RAG and 3.68 under fine-tuning (Table 2). All three pairwise differences are significant by a Wilcoxon signed-rank test paired by idiom ($+0.69$ RAG over baseline, $+0.39$ fine-tuning over RAG, $+1.08$ fine-tuning over baseline, computed from unrounded scores; each $p < 0.001$, $N = 40$), and fine-tuning exceeds RAG on every individual dimension at the same significance level. Three patterns stand out. First, fine-tuning's improvement over baseline is remarkably uniform across dimensions ($+1.05$ to $+1.10$ points each), lifting the whole quality profile rather than trading one facet for another. Second, a persistent Fluency–Adequacy gap ($+0.26$ to $+0.31$ points) holds in all three conditions: models produce well-formed isiZulu that is nonetheless less faithful to the idiom's meaning. Third, a clear gap to the ceiling remains: the baseline (2.61) sits below the scale mid-point, and even the best condition (3.68) and the best single model$\times$condition cell (3.93) stay well short of 5. The largest single-model movement is Llama 3.1 8B Instruct, whose overall rating climbs from 2.30 to 3.93 after fine-tuning.

Fine-tuning is the best condition for four of the five models; Gemma 3 12B is the exception, peaking under RAG (3.53 versus 3.25), while AfroLlama V1 has the highest baseline (2.83), narrowly ahead of Gemma 3 12B (2.82); per-model ratings are given in Appendix K. The per-model cells are not formally tested: the significance tests above are computed at condition level, pooled over the five models, and no per-model contrast is tested or corrected for multiplicity. The per-model human patterns are therefore descriptive and are read alongside the automated metrics computed on the full 323-idiom test set (Table 1).

[6] https://en.wikipedia.org/wiki/Krippendorff%27s_alpha

Table 2: Mean human ratings (1–5) by condition (three raters, five models). "Overall" is the mean of the five dimensions. The 95% confidence interval (CI) gives the likely range of each Overall mean, estimated by redrawing the 40 idioms at random 2,000 times (bootstrap). The Baseline < RAG < Fine-tuned ordering holds on every dimension, and all pairwise differences in Overall are significant (Wilcoxon signed-rank, $p < 0.001$).

Dimension	Baseline	RAG	Fine-tuned
Fluency	2.71	3.43	**3.81**
Adequacy	2.45	3.12	**3.52**
Cultural Fidelity	2.49	3.15	**3.55**
Completeness	2.68	3.38	**3.75**
Coherence	2.70	3.41	**3.79**
Overall	**2.61**	**3.30**	**3.68**
95% CI	*[2.51, 2.70]*	*[3.21, 3.39]*	*[3.59, 3.79]*

Table 3: Kendall's τ between each automated metric and the mean human rating over the 15 model×condition cells. The 95% CI gives the likely range of each τ, estimated by redrawing the 15 cells at random 10,000 times (bootstrap); * significant at $p < 0.05$. Cosine similarity is the only metric with a negative coefficient, and it is not significant.

Metric	Kendall's τ	95% CI
COMET	$+0.505^*$	$[+0.09, +0.80]$
BERTScore	$+0.498^*$	$[+0.14, +0.76]$
chrF++	$+0.314$	$[-0.09, +0.63]$
AfriCOMET	$+0.200$	$[-0.21, +0.63]$
Cosine sim.	-0.251	$[-0.63, +0.16]$

4.4 Do automated metrics agree with humans?

We compute Kendall's τ (a rank correlation that measures how often two measures order a pair of items the same way) between each metric and the mean human rating over the fifteen model×condition cells (Table 3).

Each coefficient rests on fifteen cells and carries wide uncertainty, so we report a 95% CI with each to show how much it could vary. COMET leads ($\tau = +0.505$, $p = 0.008$), closely followed by BERTScore F1 ($\tau = +0.498$, $p = 0.010$); both reward semantic fidelity over surface form. chrF++ and AfriCOMET correlate positively but not significantly. *Embedding cosine similarity is the only metric whose correlation with human judgement is negative* ($\tau = -0.251$). This coefficient is not statistically significant ($p = 0.196$; CI $[-0.63, +0.16]$) and we do not claim a proven negative correlation; its sign is, however, stable under a leave-one-cell-out check (negative in all fifteen folds, $\tau \in [-0.34, -0.18]$) and on every human dimension (Appendix L).

Because fifteen cells cannot settle the question, we add an item-level test that does not depend on the system-level correlation. Over all 1,615 scored items (323 idioms × 5 models), a Wilcoxon signed-rank test paired by idiom and model shows that fine-tuned outputs score significantly *lower* on cosine similarity than both baseline outputs (rank-biserial $r_{rb} = -0.39$) and RAG outputs ($r_{rb} = -0.54$; both Holm-corrected $p < 0.001$), although fine-tuning is the condition the experts rate highest. Cosine similarity therefore moves against human judgement on the contrast that matters most in this study. We read these results as evidence within this experiment, and as a caution against using cosine similarity unvalidated, not as a general statement about metric reliability.

4.5 Qualitative error analysis

Rater comments show a consistent progression from baseline to RAG to fine-tuning. The dominant errors are literalism, cultural flattening, truncated output, retrieval mis-integration, and hallucinated confidence (Appendix M). The most instructive case is a *RAG paradox*: on chrF++ and AfriCOMET, RAG scores *below* baseline, yet human raters place it clearly *above* baseline on every dimension. Retrieval adds correct but un-integrated glosses that pull the output away from the concise gold reference, lowering overlap scores, even as raters read them as an improvement. The paradox is thus not evidence against RAG but a further reason for caution with these metrics: the two that undervalue retrieval are ones whose correlation with human judgement was not significant.

5 Discussion

5.1 Answers to the research questions

On **RQ1**, zero-shot interpretation is far from solved: models are fluent but weak on adequacy and cultural fidelity, and baseline competence varies with the base model, with the African-pretrained model leading on the automated metrics and the general-purpose Llama models scoring lowest at baseline. On **RQ2**, retrieval helps, but only partially, and its ceiling is integration rather than coverage (recall@3 = 0.86). On **RQ3**, QLoRA fine-tuning is most effective overall, best on the human ratings and four of five automated metrics, so the conditions rank fine-tuning > RAG > baseline. On **RQ4**, no single strategy is universally best, and three as-

sociations recur. First, the weakest models gain most: the two general-purpose Llama models start lowest and gain most from fine-tuning ($+1.40$ and $+1.63$ human-rating points), although gains are not strictly ordered by baseline strength (AfroLlama V1 gains $+1.05$ from the joint-highest baseline, more than GPT-OSS 20B's $+0.87$ and Gemma 3 12B's $+0.43$). Second, the closest controlled contrast, the Llama 3.1 base/Instruct pair, behaves almost identically at baseline and under RAG, and the Instruct variant gains slightly more from fine-tuning. Third, the best strategy differs by model, and GPT-OSS 20B, the largest model, is not the highest-rated under any condition. GPT-OSS 20B, which gains least relative to its baseline (2.71), also carries a decoding caveat, since it was sampled rather than decoded greedily at baseline.

5.2 Why fine-tuning helps more than retrieval

The psycholinguistic accounts in Section 2.1 suggest an explanation. Retrieval supplies the *literal* meanings of an idiom's parts and asks the model to reason from them to the figurative meaning. This suits idioms whose meaning can be partly reconstructed from their image, as in *ikati lilele eziko* ("the cat sleeps on the hearth": no fire, so no food, so poverty). For opaque idioms, however, the figurative meaning is a conventional fact of the speech community that no chain of literal reasoning can recover; it must be *known*, as the lexical representation hypothesis holds. Fine-tuning supplies precisely this stored, whole-phrase knowledge, together with the register and structure of a good explanation, which is consistent with its uniform gain across all five human dimensions. We offer this account as a hypothesis: testing it requires grading test idioms by transparency, which we leave to future work.

The integration ceiling also has a practical face. Recall@3 is high (0.86) and every model receives the same evidence, yet GPT-OSS 20B is degraded by the same evidence that helps Gemma 3 12B. Supplying correct information is therefore necessary but not sufficient; models differ in whether they can weave glosses into a single figurative reading or merely list them. This argues for evaluating retrieval and generation separately, as we do, rather than reporting only end-to-end scores.

5.3 Implications for evaluation

Figurative generation in low-resource languages cannot be judged by surface similarity alone. We recommend reporting COMET as the primary automated proxy, treating chrF++ and AfriCOMET as complements, and not reporting embedding cosine similarity as a quality signal unless it has first been validated against human judgement. A plausible explanation for its behaviour is that it rewards broad lexical proximity to the reference, which the compact fine-tuned interpretations reduce; this would also explain why it alone rose under RAG and fell under fine-tuning. These recommendations rest on one task, one language, and fifteen systems, and should be re-validated in other settings. More broadly, the persistent Fluency–Adequacy gap shows that fluency must never be evaluated in isolation: a fluent wrong answer is the most convincing kind of error.

5.4 Implications for resourcing and heritage work

Because gains from adaptation varied widely by model, the choice of strategy should be checked against the base model at hand: in our data a weak general-purpose model gained most from fine-tuning on a modest curated corpus, whereas for models with stronger baselines the gains were smaller and the better strategy differed by model (Gemma 3 12B peaked under RAG). For communities with limited compute, this suggests that a modest, expert-verified dataset combined with parameter-efficient fine-tuning on a single GPU can close much of the gap to a model pre-trained on African languages.

For the digital humanities, the results carry a caution as well as an opportunity. Idioms such as *ucele empunzini* encode cultural knowledge that is increasingly transmitted in writing rather than speech, and LLMs could help lexicographers, teachers, and heritage projects to draft explanations and usage examples at scale. Yet the error analysis shows that even the best condition produces "hallucinated confidence": fluent, plausible explanations that are wrong. Such output should be treated as a draft for expert review, never as an authority on cultural meaning. The dataset itself (community-sourced, expert-verified, and structured as idiom, interpretation, and usage) is also a digitised heritage resource in its own right, independent of any model.

6 Conclusion

We present a fully crossed comparison of zero-shot prompting, RAG, and fine-tuning for isiZulu idiom

interpretation across five open-weight LLMs. Fine-tuning was strongest; RAG offered smaller gains that human raters perceived but surface metrics understated; and the gain from each strategy varied by base model, with the weakest models gaining most and no single strategy best for every model. We will release the dataset (where licences permit), pipeline, and evaluation protocol, and caution that similarity metrics should be validated against human judgement before they are trusted for figurative generation in low-resource languages.

Limitations

Only five models were evaluated, and they differ simultaneously in architecture, scale, pre-training data, and instruction tuning, so the model-level findings are associations rather than controlled effects; fine-tuning was limited to QLoRA and ten HPO trials per model by compute. The study covers isiZulu only: the pipeline may transfer to other Nguni Bantu languages, but this is untested, and without a parallel evaluation in a well-resourced language we cannot separate difficulties specific to isiZulu from the difficulty of idiom interpretation in general. The dataset draws substantially from a single lexicographic source, and human evaluation covers 40 idioms, a fixed convenience sample used identically across all models and conditions rather than a stratified draw; it may not fully represent the 323-item test set; per-model contrasts are untested and therefore descriptive; and the metric–human correlations rest on only fifteen system-level cells. COMET and AfriCOMET are applied off-label to monolingual generation; SSA-COMET (Li et al., 2025) and LLM-as-judge evaluation were not tested, and the latter would also inherit the judging model's limitations in isiZulu. Decoding was not fully matched: GPT-OSS 20B samples at baseline, and RAG generation samples (temperature 0.7) with a larger token budget for every model; a matched sweep is left to future work. The RAG condition bundles retrieved evidence with a chain-of-thought scaffold, worked exemplars, and English instructions, so its effect cannot be attributed to retrieval alone. Ratings were not blind to model or condition, and the raters are co-authors; the size and consistency of the effects and their agreement with the automated metrics argue against expectation bias, but an independent blind panel is the appropriate check. Finally, copyright on sources such as *Inqolobane Yesizwe* (Nyembezi and Nxumalo, 1966) constrains redistribution: code, annotations, splits, and protocols will be released where permitted, and source-derived content will follow the applicable licences.

Future work will extend evaluation to other Nguni Bantu languages and to an English idiom benchmark, combine fine-tuning with RAG, ablate the decomposition, top-k retrieval and WSD components, grade idioms by transparency, expand human evaluation so that metric–human agreement can be estimated at item level, and develop culturally aware metrics for figurative language.

Ethics Statement

Human-subjects work was approved by the Research Ethics Committee of the Faculty of Engineering, Built Environment and Information Technology (EBIT), University of Pretoria, with informed, voluntary participation and no personally identifying data. The expert raters are co-authors; the authors have no other competing interests.

Acknowledgements

The authors acknowledge the Data Science for Social Impact (DSFSI) Lab, supported by the ABSA Chair of Data Science and the African Institute of Data Science and AI (AfriDSAI). This work was conducted as part of the Artificial Intelligence for Development (AI4D) programme, with the financial support of the UK government through its Global Research and Technology Development portfolio as well as Canada's International Development Research Centre (IDRC). DSFSI also acknowledges gifts from NVIDIA, Google.org, OpenAI, and Meta. We thank the isiZulu speakers who annotated the dataset.

References

Ife Adebara and Muhammad Abdul-Mageed. 2022. Towards Afrocentric NLP for African languages: where we are and where we can go. In *Proceedings of the Association for Computational Linguistics (ACL)*.

Jesujoba Alabi et al. 2022. AfroXLMR: Adapting multilingual encoders to African languages.

Shatha Almheiri et al. 2026. Multilingual idioms in sentences and conversations across resource levels. In *Proceedings of the Association for Computational Linguistics (ACL)*.

Cristina Cacciari and Sam Glucksberg. 1991. Understanding idiomatic expressions: the contribution of

word meanings. In *Advances in Psychology*, volume 77, pages 217–240. Elsevier.

Cristina Cacciari and Patrizia Tabossi. 1988. The comprehension of idioms. *Journal of Memory and Language*.

Tim Dettmers et al. 2023. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Michael Flor, Xinyi Liu, and Anna Feldman. 2025. A survey of idiom datasets for psycholinguistic and computational research. In *Proceedings of KONVENS*.

Raymond W. Gibbs. 1980. Spilling the beans on understanding and memory for idioms. *Memory & Cognition*.

Edward J. Hu et al. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs (FAISS). *IEEE Transactions on Big Data*.

Pratik Joshi et al. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the Association for Computational Linguistics (ACL)*.

S. P. Khumalo and S. Z. Dlamini. 2025. IsiZulu idiomatic expressions derived from selected biblical figures: a morpho-syntactic & socio-cultural assimilation critique. *Cogent Arts & Humanities*, 12(1):2488651.

Jisu Kim et al. 2025. Memorization or reasoning? exploring the idiom understanding of LLMs (MIDAS). In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Klaus Krippendorff. 2004. *Content Analysis: An Introduction to Its Methodology*, 2nd edition. Sage.

Patrick Lewis et al. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Shenbin Li et al. 2025. SSA-COMET: Do LLMs outperform learned metrics in evaluating MT for under-resourced African languages? In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

C. Marquard, S. Mawere, and F. Meyer. 2025. Neural morphological tagging for Nguni languages. In *AfricaNLP Workshop*.

Maggie Mi, Aline Villavicencio, and Nafise Sadat Moosavi. 2025. Rolling the DICE on idiomaticity: how LLMs fail to grasp context. In *Proceedings of the Association for Computational Linguistics (ACL)*.

Wilhelmina Nekoto et al. 2020. Participatory research for low-resourced machine translation: a case study in African languages. In *Findings of the Association for Computational Linguistics: EMNLP*.

Geoffrey Nunberg, Ivan A. Sag, and Thomas Wasow. 1994. Idioms. *Language*.

C. L. S. Nyembezi and O. E. H. Nxumalo. 1966. *Inqolobane Yesizwe*. Shuter & Shooter, Pietermaritzburg.

Antoine Nzeyimana and Andre Niyongabo Rubungo. 2025. KinyaColBERT: a lexically grounded retrieval model for low-resource RAG. ArXiv:2507.03241.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Conference on Machine Translation (WMT)*.

Ricardo Rei et al. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Ahmed Sakhawat et al. 2026. When words don't mean what they say: figurative understanding in Bengali idioms. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*. ArXiv:2602.12921.

David A. Swinney and Anne Cutler. 1979. The access and processing of idiomatic expressions. *Journal of Verbal Learning and Verbal Behavior*.

Manish Thakur. 2025. Culturally-grounded chain-of-thought (CG-CoT) for culturally-specific tasks in low-resource languages. ArXiv:2506.01190.

Debra A. Titone and Cynthia M. Connine. 1999. On the compositional and noncompositional nature of idiomatic expressions. *Journal of Pragmatics*.

N. Tsotetsi, L. van Huyssteen, and M. L. Kabinde Machate. 2025. Locatives in isiZulu: a grammatico-linguistic analysis of usage trends. *South African Journal of African Languages*, 45(1):20–26.

Jiayi Wang et al. 2024. AfriMTE and AfriCOMET: enhancing COMET to embrace under-resourced African languages. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Suhang Wu et al. 2024. Not all languages are equal: insights into multilingual retrieval-augmented generation. ArXiv:2410.21970.

Tianyi Zhang et al. 2020. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations (ICLR)*.

Appendix Overview

The appendices follow the paper's own structure. Appendix A collects abbreviations and technical terms. Appendix B documents the **dataset**. Appendices C–G cover the **methods and model configuration** in the order the conditions appear in Section 3: the baseline and fine-tuned prompt templates (Appendix C), the supervised fine-tuning data format (Appendix D) and configuration (Appendix E), and the RAG pipeline's morphological pre-processing (Appendix F) and prompt templates (Appendix G). Appendices H–L give the **evaluation details**: the human-evaluation rubric, inter-rater agreement, dimension separability, per-model human ratings, and per-dimension metric–human agreement. Appendix M presents **representative model outputs**.

A Glossary of Terms and Abbreviations

Table 4 spells out every abbreviation and technical term used in the paper.

B Dataset Sources and Construction Details

The isiZulu Idiom Dataset draws on both digital and print lexicographic resources, selected for their established community grounding and lexicographic authority.

Digital sources. The digital sources consulted include dedicated isiZulu language and cultural websites (*Zuluring Blog*,[7] *Indabukoyakho*,[8] *Izithakazelo Blog*,[9] *iAfrika*,[10] *Emandulo Kingdom*,[11] and the *Ulwazi Programme*[12]), as well as isiZulu Wikipedia and general isiZulu-language social media content.

Print source. A primary print lexicographic resource was also used: *Inqolobane Yesizwe* by C. L. S. Nyembezi and O. E. H. Nxumalo, published in 1966 by Shuter and Shooter (Nyembezi and Nxumalo, 1966).

Construction and verification. The structure of each entry, the range of idioms covered, and the expert verification procedure are described in Section 3.2.

[7] https://zuluring.blogspot.com
[8] https://indabukoyakho.com
[9] https://izithakazelo.blog
[10] https://iafrika.org
[11] https://emandulokingdom.wordpress.com
[12] https://www.ulwaziprogramme.org

Table 4: Glossary of abbreviations and technical terms. In the PDF, each linked term (in colour) opens an introductory explanation.

Term	Meaning
LLM	Large language model: a neural network trained on large text collections to predict and generate text.
NLP	Natural language processing: computational analysis and generation of human language.
Zero-shot	Asking a model to perform a task with no examples or task-specific training.
RAG	Retrieval-augmented generation: looking up relevant documents and adding them to the prompt before the model answers.
Fine-tuning	Further training of a pre-trained model on task-specific examples.
LoRA / QLoRA	(Quantised) Low-Rank Adaptation: fine-tuning small add-on matrices instead of all weights; QLoRA also stores the backbone in 4-bit precision.
NF4	4-bit NormalFloat, the compressed number format used by QLoRA.
PEFT	Parameter-efficient fine-tuning; also the name of the Hugging Face library used.
Unsloth	An open-source library for faster, lower-memory LoRA/QLoRA fine-tuning.
BitsAndBytes	The library that provides 4-bit (NF4) quantisation for QLoRA.
FastText	A word-embedding method that represents words by character n-grams, so related word forms get similar vectors.
JSON	JavaScript Object Notation: a plain-text format for structured data.
HPO / TPE	Hyperparameter optimisation; tree-structured Parzen estimator, the Bayesian search method in Optuna.
GPU	Graphics processing unit, the hardware used to train and run models.
FAISS	Facebook AI Similarity Search: a library for fast search over vector embeddings.
Embedding	A numeric vector representing the meaning of a word or text; similar meanings lie close together.
Cosine similarity	Similarity between two embeddings, measured by the angle between them.
WSD	Word-sense disambiguation: choosing which meaning of a polysemous word is intended in context.
Chain-of-thought	A prompt asking the model to reason step by step before answering.
Recall@k / MRR	Share of words whose correct entry is in the top k results; mean reciprocal rank of the correct entry.
BLEU	Bilingual Evaluation Understudy: a word n-gram overlap metric.
chrF++	Character and word n-gram F-score (Popović, 2017).
BERTScore	Similarity of contextual token embeddings of output and reference (Zhang et al., 2020).
COMET	Crosslingual Optimized Metric for Evaluation of Translation, a metric trained on human ratings (Rei et al., 2020).
AfriCOMET	COMET retrained for African languages (Wang et al., 2024).
SSA-COMET	Sub-Saharan Africa COMET (Li et al., 2025).
DA	Direct Assessment: raters score each output on an absolute scale.
Krippendorff's α	Chance-corrected inter-rater agreement coefficient.
Kendall's τ / Spearman's ρ	Rank correlation coefficients.
Wilcoxon test	Wilcoxon signed-rank test: a non-parametric test for paired differences.
CI	Confidence interval.
RQ	Research question.

C Baseline and Fine-Tuning Prompts

This appendix reproduces, verbatim, the prompt templates used in the baseline (zero-shot) and fine-tuned inference conditions described in Section 3. In every template the placeholder `{idiom}` is replaced at run time with the idiom under evaluation; all other elements are held constant across idioms and models.

Prompt configuration. In the baseline condition the prompt is the sole mechanism through which task context is communicated to the model; it is therefore the critical design variable distinguishing the baseline from the other conditions. The required output is a standardised three-part scaffold in isiZulu: an *Isisho* section reproducing or labelling the input idiom, an *Incazelo* section providing a concise semantic interpretation in isiZulu, and an *Izibonelo* section comprising numbered contextual example sentences. This scaffold mirrors the structure used in the fine-tuning and retrieval-augmented conditions, ensuring that differences in output quality between conditions are attributable to the experimental manipulation rather than to formatting differences.

The default zero-shot prompt (Box 1) was applied identically to all five models. The prompt opens with the isiZulu instruction *Chaza lesisisho sesiZulu* ("Explain this isiZulu idiom"), directs the model to answer using one fixed format (*Phendula ngale fomethi kuphela*, "Answer using this format only"), and then presents the *Isisho*/*Incazelo*/*Izibonelo* template, which the model completes. No worked examples are embedded in this prompt, so it is strictly zero-shot. The complete template is shown in Box 1.

Box 1: Uniform zero-shot baseline prompt (applied identically to all five models)

```
Chaza lesisisho sesiZulu: "{idiom}"
Phendula ngale fomethi kuphela:
Isisho: {idiom}
Incazelo: [chaza ngesiZulu]
Izibonelo:
1. [umusho wokuqala ngesiZulu]
2. [umusho wesibili ngesiZulu]
3. [umusho wesithathu ngesiZulu]
Isisho:
```

Box 2 gives the fine-tuned inference prompt. Because the *Isisho*/*Incazelo*/*Izibonelo* scaffold is learned during fine-tuning (Appendix D), the fine-tuned models require no format instructions at inference time, so the prompt reduces to the bare instruction and the target idiom. The model reproduces the three-part structure from its learned behaviour rather than from the prompt, so any residual format errors are attributable to training rather than to prompt under-specification.

Box 2: Fine-tuned inference prompt (format learned during training, not prompted)

```
Chaza lesisisho sesiZulu: "{idiom}"
```

D Supervised Fine-Tuning Data Format

The raw JSON records were converted to a supervised fine-tuning format suitable for training a causal language model under an instruction-following paradigm. Each record was transformed into a single string encoding both the prompt and the target completion using the model's native special tokens, following the structure in Box 3 (shown with Llama's special tokens; the other models use their own equivalents).

Box 3: Supervised fine-tuning string format (native special tokens)

```
<|begin_of_text|>Chaza lesisisho
sesiZulu: "idiom"
Isisho: idiom
Incazelo: explanation
Izibonelo: 1. example_1
2. example_2
3. example_3<|end_of_text|>
```

E Fine-Tuning Configuration and HPO Search Space

This appendix gives the QLoRA fine-tuning configuration and the shared hyperparameter search space summarised in Section 3.

LoRA adapters. LoRA decomposes each weight update into a low-rank product BA ($r \ll d, k$), so $W' = W + BA$ with W frozen and only A, B trained; at $r = 16$ a single 4096×4096 projection falls from ≈16.7M to ≈131K trainable parameters (∼128:1). A fixed adapter configuration (rank and α) is applied uniformly across all five models; adapter dropout is the only setting that differs between the PEFT and Unsloth training pipelines (see below).

Target modules. The decomposition is applied to all seven linear projections per transformer block: the four self-attention projections and the three gated feed-forward projections.

NF4 quantisation. Each backbone is loaded in 4-bit NormalFloat (NF4) via BitsAndBytes while the adapters train in `bfloat16`. Only adapter-loading dropout differs between the PEFT and Unsloth pipelines; all other settings are held constant.

Hyperparameter optimisation. Bayesian hyperparameter optimisation using a tree-structured Parzen estimator (TPE, via Optuna) minimises loss on the 323-idiom validation set over a fixed budget of ten trials per model, five epochs each, within the shared search space in Table 5. The single-GPU constraint that motivates QLoRA also bounds this budget, particularly for the 20-billion-parameter model.

Table 5: Shared hyperparameter search space applied uniformly across all five models in the Optuna TPE optimisation (ten trials each).

Hyperparameter	Strategy	Range
Learning rate	log-uniform	[5e-5, 3e-4]
Warmup ratio	uniform	[0.01, 0.10]
Grad. accumulation	categorical	$\{1, 2, 4\}$
LR scheduler	categorical	linear / cosine

F Morphological Pre-processing

IsiZulu exhibits rich agglutinative morphology, with noun-class prefixes (*umu-*, *isi-*, *ama-*, *aba-*, *ili-*, *in-*, *iz-*, *u-*, *i-*), locative forms (*e-…-ini*), and verbal inflections (*-ile*, *-ela*, *-isa*) that regularly obscure the relationship between a surface form and its dictionary lemma (Marquard et al., 2025; Tsotetsi et al., 2025). The RAG pipeline implements a rule-based morphological decomposition module applying systematic prefix-stripping and suffix-stripping transformations, generating candidate root forms for each word.

The decomposition rules are organised into four categories: locative prefix and suffix stripping (e.g. *e-…-ini* → *i-…*), tense-marker removal, verbal-suffix removal (*-ile*, *-ela*, *-isa*, *-ana*, *-eka*, *-wa*, *-iswa*, *-elwa*, *-iwe*), and noun-class prefix stripping. The noun-prefix rules additionally encode phonological coalescence (for example, when the prefix *a* precedes a stem beginning with *i*, the sequence contracts to *e*) and plural-to-singular alternations, including *izi-* to *isi-*, *ama-* to *ili-* and *i-*, and *aba-* to *umu-*, *um-*, and *u-*. An isiZulu phonological validity filter (`is_valid_zulu_form`) discards candidate roots shorter than three characters or lacking a vowel. This approach is acknowledged as a rule-based approximation rather than a full morphological analyser, consistent with the position advanced by Marquard et al. (2025) that for low-resource Bantu languages, rule-based approximations often achieve comparable practical recall at substantially lower development cost.

G RAG Prompt Templates and Worked Examples

This appendix reproduces, verbatim, the two prompt templates used in the RAG pipeline: the word-sense-disambiguation (WSD) prompt (Box 4) and the chain-of-thought explanation prompt (Box 5), each of which embeds worked examples.

Box 4: WSD prompt, system instructions and rules (verbatim)

```
You are an isiZulu linguist performing WORD
SENSE DISAMBIGUATION for idioms.
TASK: For each word in the idiom below, you are
given candidate glossary definitions.  For each
word, do ONE of the following:
SELECT - Pick the candidate that best fits the
idiom context.
REJECT - If NONE of the candidates are correct,
reject them all and provide the correct meaning
yourself.  This happens when the glossary
matching returned wrong words (e.g., matched
"umxopho" (marshy place) when the actual word is
"umopho" (blood)).
RULES:
1.  Think about how the words work TOGETHER to
form a coherent literal image.
2.  Idioms paint a PICTURE. Choose meanings that
create a scene, not abstract concepts.
3.  Consider isiZulu morphology:  the surface
form may differ from the glossary lemma (e.g.,
lilele is perfect tense of lala; eziko is
locative of iziko).
4.  If a candidate's base lemma does not match
the actual root of the word, REJECT it.
5.  When rejecting, provide the correct meaning
based on your knowledge of isiZulu.
6.  Output ONLY valid JSON.
[THREE WORKED EXAMPLES FOLLOW]
NOW DISAMBIGUATE:
Idiom:  "{idiom}"
Candidates:
{word_1}:
  1.  ({lemma}) definition [score=X.XX]
  2.  ({lemma}) definition [score=X.XX]
  3.  ({lemma}) definition [score=X.XX]
{word_2}:  ...
For each word, output JSON with action SELECT or
REJECT:
```

Box 5: RAG chain-of-thought explanation prompt (verbatim)

```
You are an expert isiZulu linguist specializing
in idioms and figurative language.
TASK: Given an isiZulu idiom and glossary
definitions of its component words, reason
step-by-step from the LITERAL word meanings to
the FIGURATIVE meaning.
PROCESS (follow this exactly):
1.  GLOSSARY LOOKUP: Note what each word means
literally from the definitions provided.
2.  LITERAL IMAGE: What scene or picture do
these literal meanings create together?
3.  FIGURATIVE REASONING: WHY would people use
this literal image to describe a situation?
```

```
What is the logical chain from the literal scene
to the figurative meaning?
4.  STATE THE MEANING: What does the idiom
actually mean, concisely, in isiZulu?
5.  GENERATE EXAMPLES: Write 5 natural isiZulu
sentences using the idiom in different contexts.
RULES: The "explanation" field MUST be in
isiZulu.  All example sentences MUST be in
isiZulu.  Your reasoning must be a clear CHAIN:
word meanings → literal scene → why figurative
→ meaning.  Do NOT rely on prior knowledge of
the idiom.  Deduce the meaning ONLY from the
glossary.  After reasoning, output ONLY a valid
JSON object.
=== WORKED EXAMPLE (EXAMPLE 1:  ikati lilele
eziko) ===
Step 1 - Glossary Lookup:
  ikati = isilwane esincane esifuyiwe (cat)
  lala (lilele) = ukubeka umzimba phansi ulale
ubuthongo (to sleep)
  iziko (eziko) = indawo exhibeni lapho
kubaselwa khona (hearth)
Step 2 - Reasoning:
  Literal image:  A cat is sleeping at the
cooking hearth.
  Figurative reasoning:  A cat sleeps at the
hearth only if there is no fire.  No fire means
no one is cooking; no cooking means no food;
poverty.
  Meaning:  Kunenhlupheko / Kunokushoda kokudla
endlini.
Step 3 - Final Output (JSON): { "idiom":  "ikati
lilele eziko", "explanation":  "Kushiwo uma
kuhluphekwa ekhaya...", "examples":  [...]  }
[EXAMPLES 2 AND 3 FOLLOW THE SAME STRUCTURE]
=== NOW EXPLAIN THIS IDIOM ===
IDIOM: "{idiom}"
{glossary_context_block}
INSTRUCTIONS: (1) First, write your reasoning
(Steps 1-3 as shown).  (2) Then output the JSON
object with fields idiom, explanation (isiZulu,
concise), examples (5 isiZulu sentences),
and reasoning (word_meanings, literal_image,
figurative_reasoning).  Begin:
```

H Human Evaluation Rubric

Table 6 defines the five Direct Assessment dimensions, each scored 1–5.

Table 6: The five human-evaluation dimensions.

Dimension	Facet of quality
Fluency	Grammatical, natural isiZulu, independent of correctness.
Adequacy	Semantic faithfulness to the idiom's true meaning.
Cultural Fidelity	Correctness of cultural and pragmatic usage.
Completeness	Coverage of expected components and examples.
Coherence	Logical structure and internal consistency.

I Inter-Rater Agreement: Krippendorff's α

Reliability is quantified with Krippendorff's α (Krippendorff, 2004), defined as

$$\alpha = 1 - \frac{D_o}{D_e}, \tag{1}$$

where D_o is the observed disagreement and D_e the disagreement expected by chance,

$$\begin{aligned} D_o &= \frac{1}{n} \sum_c \sum_k o_{ck}\, \delta^2(c,k), \\ D_e &= \frac{1}{n(n-1)} \sum_c \sum_k n_c\, n_k\, \delta^2(c,k), \end{aligned} \tag{2}$$

with o_{ck} the number of rater pairs assigning values c and k to the same item, n_c the total number of c ratings, n the number of pairable values, and the interval distance $\delta^2(c,k) = (c-k)^2$, which treats the 1–5 rating scale as interval data.

J Dimension Separability

Table 7 reports pairwise Spearman correlations among the five dimensions over all 1,800 ratings. Values are strong but well below 1, so raters did not collapse the dimensions into a single quality axis. Fluency–Adequacy ($\rho = 0.77$), the two facets that the study finds diverge most often, is among the least-correlated pairs.

Table 7: Spearman correlations among the human-evaluation dimensions.

	Flu.	Adeq.	Cult.	Comp.	Coh.
Fluency	1.00	0.77	0.78	0.82	0.86
Adequacy	0.77	1.00	0.84	0.76	0.75
Cultural Fid.	0.78	0.84	1.00	0.79	0.78
Completeness	0.82	0.76	0.79	1.00	0.83
Coherence	0.86	0.75	0.78	0.83	1.00

K Per-Model Human Ratings

Table 8 gives the mean overall human rating for each model under each condition (120 ratings per cell), with each model's gain from RAG and from fine-tuning relative to its own baseline.

Table 8: Mean overall human rating (1–5) per model and condition. Best per model in **bold**. Gains are computed from unrounded means, so they can differ by 0.01 from the printed values.

Model	B	R	F	ΔR	ΔF
AfroLlama V1	2.83	3.14	**3.88**	+0.31	+1.05
Llama 3.1 8B	2.36	3.30	**3.77**	+0.94	+1.40
Llama 3.1 8B-It	2.30	3.22	**3.93**	+0.92	+1.63
GPT-OSS 20B	2.71	3.29	**3.58**	+0.58	+0.87
Gemma 3 12B	2.82	**3.53**	3.25	+0.71	+0.43
Mean	2.61	3.30	**3.68**	+0.69	+1.08

L Per-Dimension Metric–Human Agreement

Table 9 extends Table 3 to each human-evaluation dimension, again computed over the 15 model×condition cells. The ordering of metrics is stable across dimensions: COMET and BERTScore F1 correlate positively and significantly on every dimension, chrF++ and AfriCOMET positively but not significantly, and cosine similarity negatively on every dimension, though none of its coefficients is statistically significant.

Table 9: Kendall's τ between each automated metric and the mean human rating, overall and per dimension, over the 15 model×condition cells (Flu. = Fluency, Adeq. = Adequacy, Cult. = Cultural Fidelity, Comp. = Completeness, Coh. = Coherence; * significant at $p < 0.05$).

Metric	Overall	Flu.	Adeq.	Cult.	Comp.	Coh.
COMET	$+0.505^*$	$+0.54^*$	$+0.50^*$	$+0.56^*$	$+0.48^*$	$+0.49^*$
BERTScore	$+0.498^*$	$+0.53^*$	$+0.42^*$	$+0.47^*$	$+0.53^*$	$+0.56^*$
chrF++	$+0.314$	$+0.34$	$+0.28$	$+0.29$	$+0.34$	$+0.37$
AfriCOMET	$+0.200$	$+0.21$	$+0.24$	$+0.25$	$+0.13$	$+0.14$
Cosine sim.	-0.251	-0.24	-0.25	-0.24	-0.20	-0.27

M Representative Model Outputs

This appendix presents three representative system outputs spanning the conditions and models, together with the mean score awarded by the expert panel. They illustrate the qualitative progression from baseline to RAG to fine-tuning discussed in Section 4.

Box 6: RAG output from Gemma 3 12B (human mean 4.0/5)

Idiom: *Ukuqhumisa uhleko*
Reference: *Ukwephula uhleko; ukuhlekela phezulu*
Incazelo: *Ukuqala ukuhleka ngamandla kakhulu kuye kwaba ngokushesha.*
Izibonelo:
1. *Lapho esezwa indaba, waqala ukuqhumisa uhleko.*
2. *Imfihlakalo yakhe yaqhumisa uhleko lwayo.*
3. *Abazali bakhe baqhumisa uhleko ngesikhathi eseyidlula iminyaka.*
4. *Ngemuva kokuba wayeqala ukuqhumisa uhleko, wanquma ukuthi ayohamba.*
5. *Ukuqhumisa uhleko kwakhe kwavusa injabulo emndenini.*

WSD decisions (glossary sense selected per word):
ukuqhumisa → **qhumuka**: "Burst into a fit of laughter" (chosen over "start off as oxen for home/water"; the idiom describes a fit of laughter, so the laughter sense fits the context).
uhleko → **hleka**: "Laugh; laugh at" (direct sense of the noun *uhleko*).

Reasoning (literal image): Someone bursting into a fit of laughter.

Figurative reasoning: *qhumisa* (burst) suggests a sudden, uncontrollable release; combined with *uhleko* (laughter) it evokes intense, explosive laughter, signifying a situation that is extremely funny.

Box 7: Baseline output from AfroLlama V1 (human mean 2.6/5): literalism

Idiom: *Uyosinda empini*
Reference: *Ufike ukudla obekade kudliwa sekuphelile*
Response: *Lesi sisho sisetshenziselwa ukubonisa ukuthi umuntu uzokwazi ukunqoba empini.*
Izibonelo:
1. *Ngikholwa ukuthi ngizokwazi ukunqoba le mpi.*
2. *Kufanele uhlale unolaka futhi uqine ukuze usinde empini.*

Box 8: Fine-tuned output from Llama 3.1 8B (human mean 4.0/5)

Idiom: *Ukuwuhlaba uwulawulele*
Reference: *Ukuhlaba umkhosi*
Isisho: *Ukuwuhlaba uwulawulele*
Incazelo: *Ukumemeza; ukuxwayisa*
Izibonelo:
1. *Umfundisi wathi, "Sengiwuhlabe uwulawulele wonke umuntu ophuma endlini engavumelekile."*
2. *UNokuthula wahlabe uwulawulele ukuze axwayise abangane bakhe.*
3. *Abangane baxoxa ukuthi ukuwuhlaba uwulawulele kudinga isibindi.*