

When Audit Quality Fails to Predict Downstream Utility

Synthetic-Data Selection for Low-Resource Classification

Xuan-Bach Le^{1†}, Phat T. Tran-Truong^{1*}, Son Ha Xuan²

¹Ho Chi Minh City University of Technology (HCMUT), VNU-HCM ²RMIT University, Vietnam [†]Equal contribution ^{*}Corresponding author
lexuanbach@hcmut.edu.vn phatttt@hcmut.edu.vn ha.son@rmit.edu.vn

EMNLP 2026



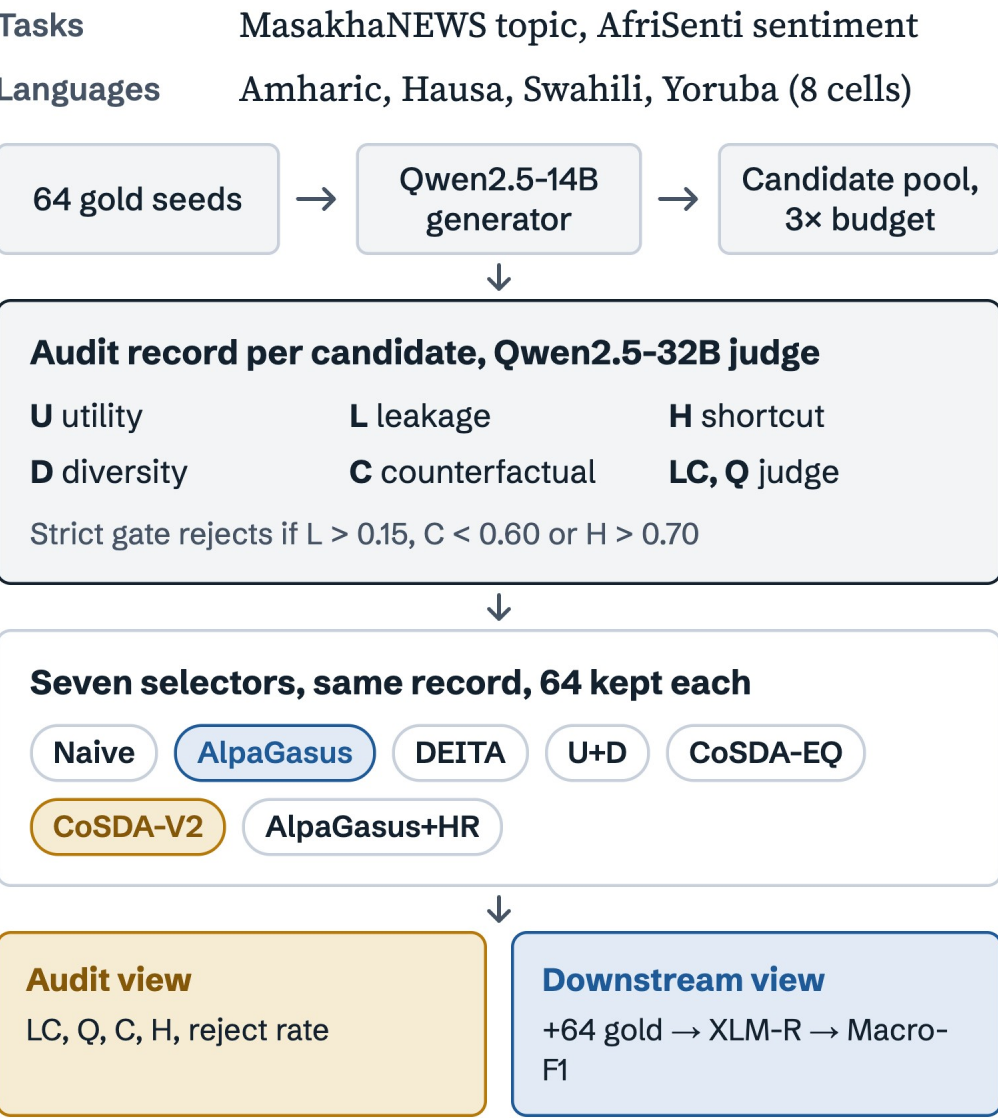
github.com/lexuanbach/CoSDA
Code, audit records, retained pools

CoSDA-V2 earns the best audit scores (label correctness 0.904), yet **AlpaGasus** trains the best classifier (Macro-F1 0.202). Inside each cell, audit rank and downstream rank barely agree: mean Spearman ρ = +0.04.

Question

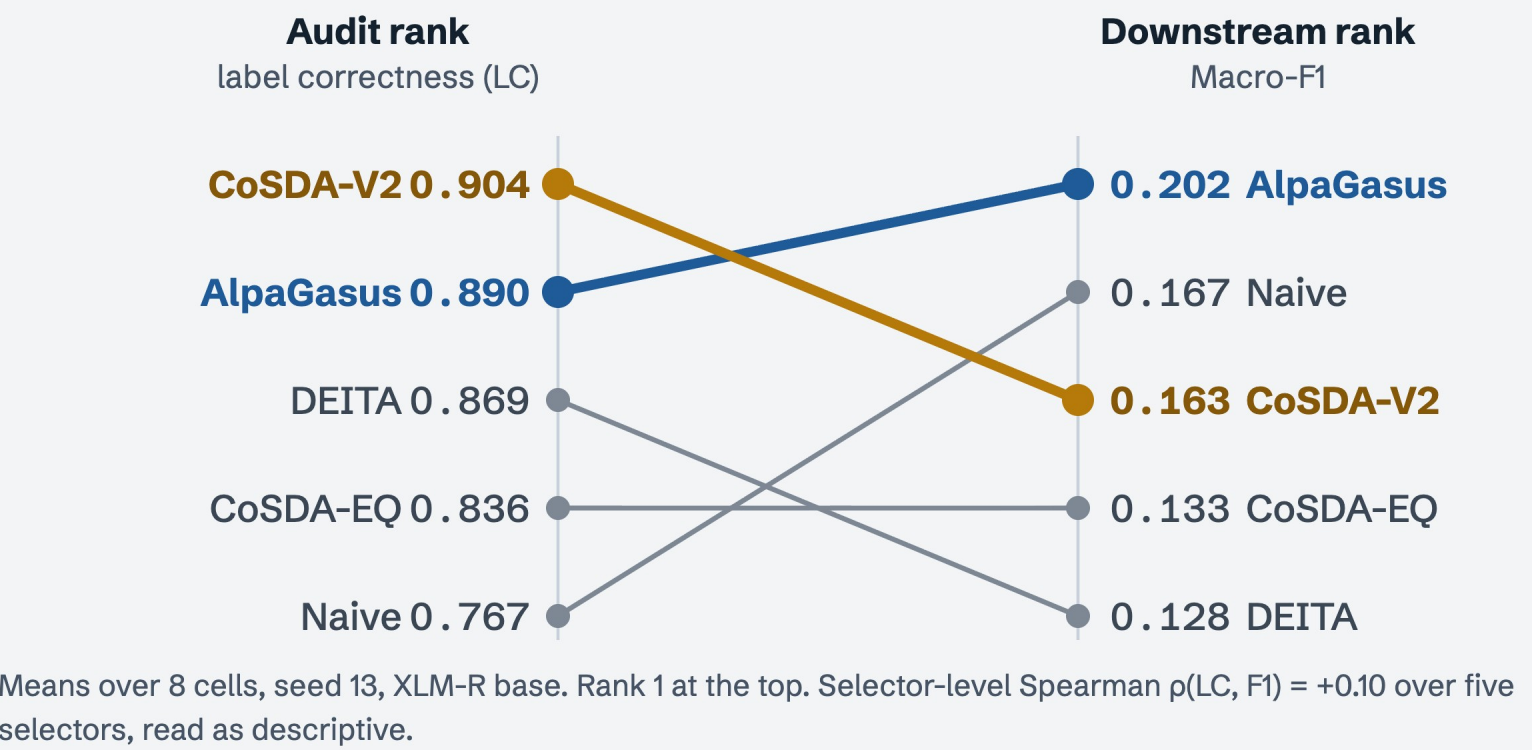
Quality-aware selectors keep the synthetic examples an LLM judge rates highest. Do audit scores predict downstream utility at a matched budget?

Controlled replay

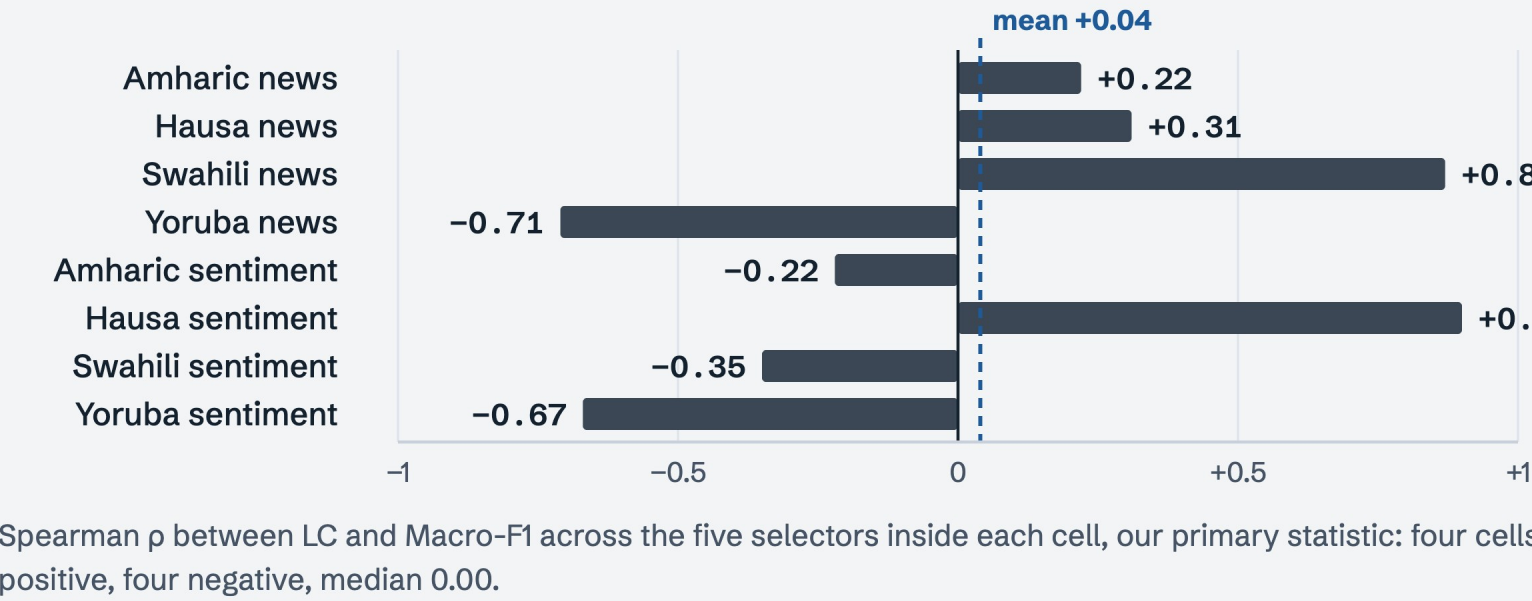


Audit and downstream rankings disagree

Every quality-aware selector beats naive on LC, Q, C and reject rate. Downstream, only AlpaGasus beats naive.



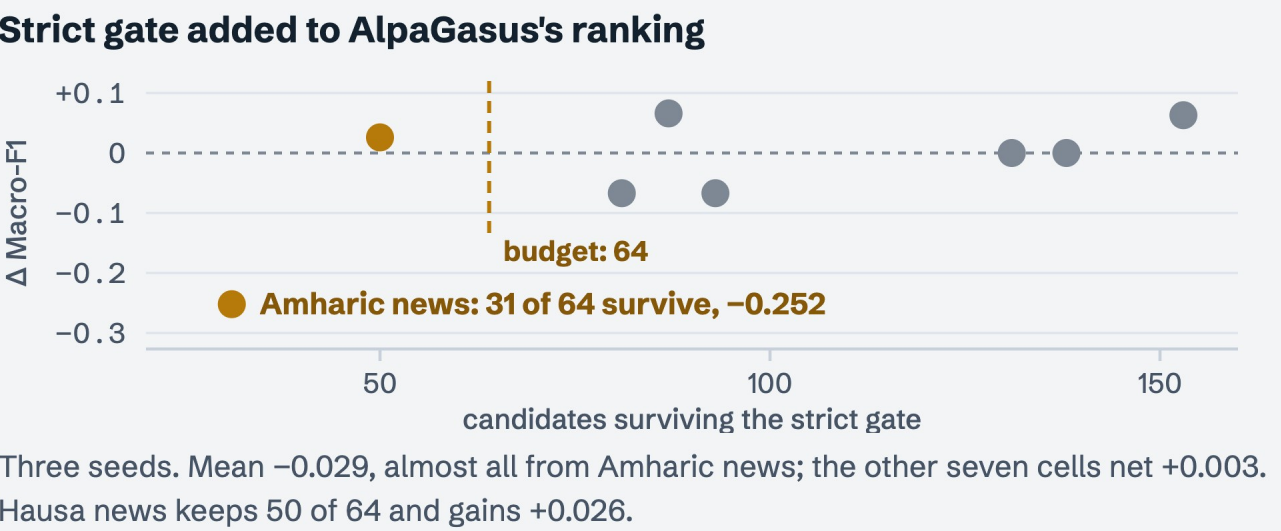
Within each cell, agreement splits evenly



The gap persists under stress

Seeds 13, 21, 42	CoSDA-V2 falls from 2nd to 6th to last of 7. Selector-level $\rho(\text{LC}, \text{F1})$ runs +0.46, +0.43, -0.29.
Backbone AfroXLMR base	All selectors gain; within-cell ρ rises to +0.21 and CoSDA-V2 moves to 2nd.
Judges Qwen-14B, Phi-3.5	LC correlates +0.73 and +0.29 with the main judge, and the LC leader changes.

Retention constraints outweigh ranking



LESS: label balance decides

U+D control	0.297
AlpaGasus	0.288
LESS with quotas	0.278
Naive, no quotas	0.205
LESS, plain top-k	0.138

Real LESS, seed 13. Plain top-k puts 77% of the pool in one label. Quotas add +0.140; reranking under quotas moves it by at most 0.019.

Takeaways

Report both metrics on the same retained sets

Audit and downstream rankings can diverge even when generation is held fixed.

Treat judged LC as screening, not a utility proxy

Selector-level $\rho(\text{LC}, \text{F1})$ swings from -0.29 to +0.46 across seeds.

Make hard rejection budget-aware

The gate's cost concentrates where too few candidates survive.