Evolving LLM-Generated Features for Interpretable Classification

Jack Butler, Zainab Afolabi, Nikita Kozodoi
AWS Forward Deployed Engineering

amazon | science

1 Problem

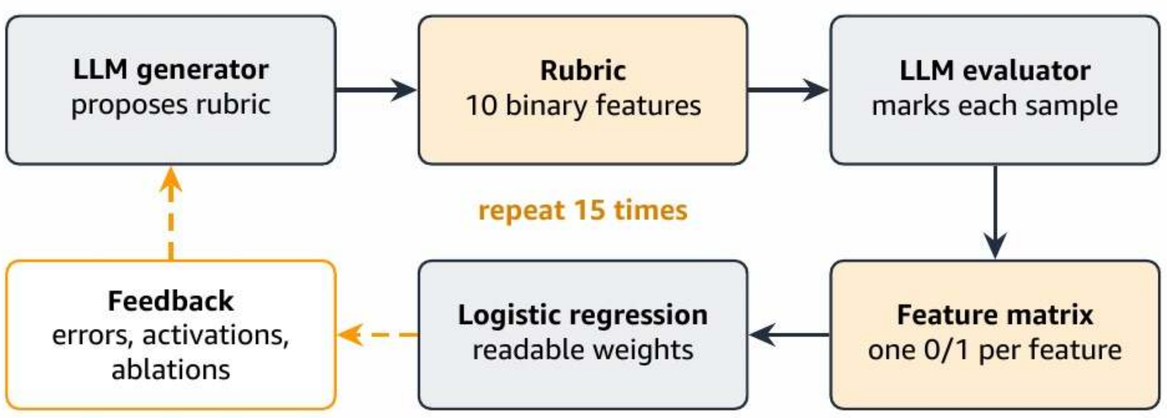
LLMs are strong classifiers, but their decisions are **opaque**. In regulated settings such as credit scoring, each decision needs a reason a person can check and dispute.

Aggregate accuracy can also hide a broken classifier: a model stuck on **one class** looks the same as a model facing a hard task.

Key Findings

- Evolved rubrics beat single-shot rubrics by **+2.9 pp** on average.
- They beat zero-shot LLM classification on **2 of 3** tasks, with every decision readable.
- On HELOC credit risk the zero-shot LLM is at chance (50.7%) and predicts `at_risk` for **81%** of a balanced test set. Only a per-class audit shows it.

2 Evolving Rubrics



Each iteration shows the generator its **confused class pairs**, each feature's **per-class activation rate**, and **ablation scores** (essential, useful or unnecessary). It keeps the strongest features and replaces the rest. A new rubric is accepted only if 5-fold CV accuracy improves.

3 Setup

- **Datasets:** AG News (4 news topics), GoEmotions (7 emotions), HELOC (binary credit risk, tabular data as text).
- **Claude Haiku 4.5** as generator and evaluator. 500 samples per dataset, 80/20 split, 3 seeds.
- **Baselines:** TF-IDF with logistic regression, **zero-shot features** (the first rubric, no feedback), and the **zero-shot LLM** labelling directly.
- **Cost:** about 5,000 LLM calls to train. Inference is one call per sample, the same as the zero-shot LLM.

4 The Rubric Is the Audit Trail

On HELOC, evolution finds numerical thresholds: decision stumps written in plain language.

Top 5 of 10 evolved features, HELOC	|w|
Net Fraction Of Revolving Burden > 40	0.65
Number Of Satisfactory Trades < 15	0.54
Months Since Last Illegal Trade between 0 and 12	0.51
Percentage Of Legal Trades < 90	0.40
Months Since First Trade < 100 and ≥ 3 trades opened in the last year	0.37

A decision then reads: *at risk because revolving burden > 40 and ≥ 1 trade insolvent for over 60 days*. Each line can be checked against a credit report.

5 Results

	TF-IDF	Zero-shot features	Evolved features	Zero-shot LLM
AG News	80.0±4.1	82.7±0.5	81.3±2.6	**89.7**±1.2
GoEmotions	31.3±8.2	34.7±6.2	**41.0**±1.6	38.7±3.7
HELOC	62.3±6.0	65.3±5.7	**69.0**±2.2	50.7±1.7
Average	57.9	60.9	**63.8**	59.7

Table 1. Test accuracy (%), mean ± std over 3 splits.

- Largest gains over zero-shot features: **GoEmotions +6.3 pp**, **HELOC +3.7 pp**.
- On AG News the labels explain themselves: the zero-shot LLM wins, and evolution does not improve on the first rubric.

6 Accuracy Hides a Failure

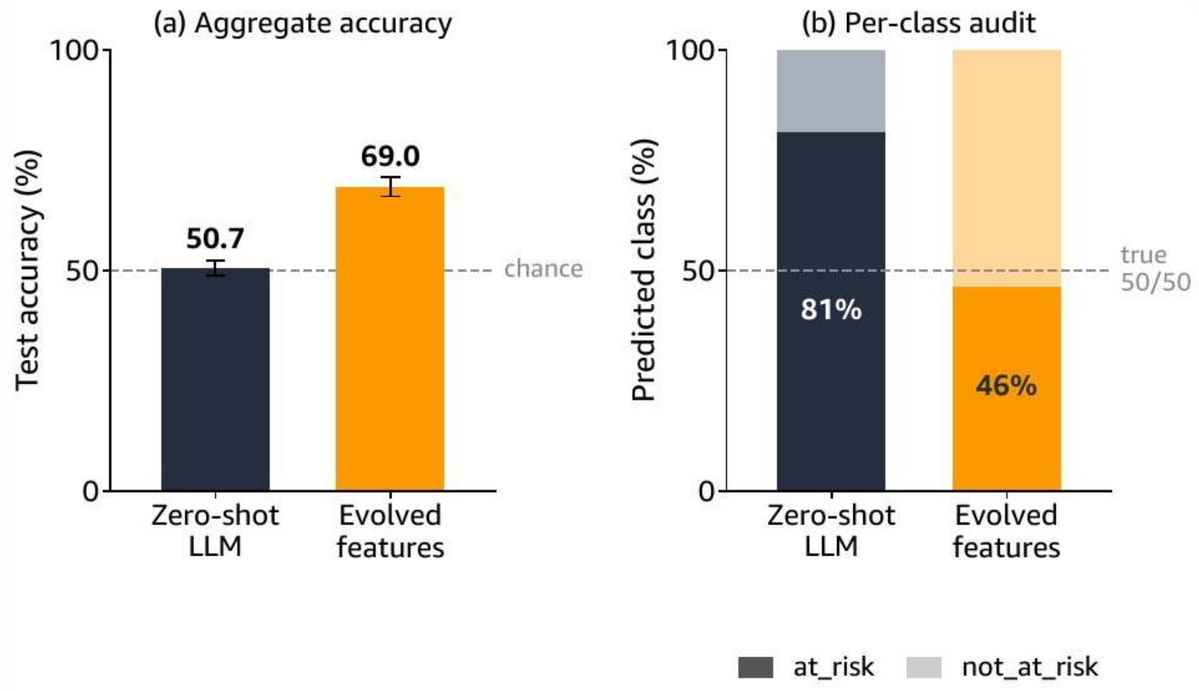


Figure 1. HELOC. (a) 50.7% reads as a hard task. (b) On a 50/50 test set the zero-shot LLM predicts `at_risk` 81% of the time, with `not_at_risk` recall of only 0.14 to 0.26. Evolved features stay balanced, with per-class F1 of 0.64 to 0.71 on both classes.

7 Take-aways

For evaluation practice

- Report **per-class metrics** before trusting an aggregate score.
- An interpretable feature layer makes that audit routine.
- The feature extractor is an LLM judge: validate it separately.
- Evolution helps most when labels are not self-explanatory or the LLM lacks domain reasoning.

Paper

arxiv.org/pdf/2610.03951

jackbtlr@amazon.com