AI4WINE: A Generative Machine Learning Benchmark for Wine Quality Classification

Valentin Pletea-Marinescu
Automation and Industrial Informatics
University Politehnica of Bucharest
Bucharest, Romania
valentin.pletea@stud.acs.upb.ro

Mihai-Daniel Pavel
Automation and Industrial Informatics
University Politehnica of Bucharest
Bucharest, Romania
mihai_daniel.pavel@upb.ro

Grigore Stamatescu
Automation and Industrial Informatics
University Politehnica of Bucharest
Bucharest, Romania
grigore.stamatescu@upb.ro

***Abstract*—Evaluating wine quality from expert review text is a natural language processing task with increasing applied value, yet retrieval-augmented prompting of open-weights large language models (LLMs) has been understudied in this context. We benchmark three instruction-tuned LLMs (Mistral-Nemo-12B, Gemma-3-12B, Qwen-2.5-14B) across zero-shot, random few-shot, and retrieval-augmented prompting with class-balanced and global exemplar selection, using frozen sentence embeddings and a corpus-informed system prompt whose statistical priors are derived exclusively from the training split. On a stratified 500-review subsample of the Wine Reviews Ordinal test split, evaluated with bootstrap confidence intervals and exact McNemar tests, the best configuration (Gemma-3-12B with global retrieval) attains 0.868 accuracy (95% CI [0.838, 0.896]) and the highest macro-F1 (0.744) of all evaluated systems, while retrieval augmentation improves Mistral-Nemo-12B by 8.0 percentage points over zero-shot ($p < 0.001$). With zero gradient updates, the strongest prompted LLMs reach statistical parity with a TF-IDF logistic-regression baseline trained on the full 71.5k-review training split. These results support retrieval-augmented in-context learning as a practical alternative to supervised training for ordinal review classification on commodity hardware.**

***Index Terms*—Large language models, Wine industry, Retrieval-augmented generation, Natural language processing, Text classification**

I. Introduction

Accelerated by the pandemic and the globalization of e-commerce, the wine sector is one of the fastest adopters of digital transformation, and wine quality assessment has been identified as a promising research field bridging oenology and artificial intelligence [1]. Past attempts parse product attributes, such as physicochemical properties like acidity, pH, density or alcohol content, into features classified with classical machine learning algorithms, such as Random Forest, SVM or Naive Bayes [2]–[4]. While this quantitative approach offers precision and objective consistency [5], it overlooks the qualitative, sensorial dimension of wine, captured mostly in reviews written by critics and sommeliers [6].

In this paper we present an enhanced machine learning pipeline based on the deployment of open-weights instruction-tuned LLMs (Mistral-Nemo, Gemma-3, Qwen-2.5) for wine quality classification. Building on Retrieval-Augmented Generation [7] and few-shot prompt engineering [8], the pipeline retrieves the k most similar labelled examples via frozen semantic sentence embeddings [9] and requires no fine-tuning. Beyond the pipeline itself, the contributions are: (i) a controlled ablation isolating the effect of class-balanced versus global exemplar retrieval; (ii) a corpus-informed system prompt whose statistical priors (price, review length, grape variety, origin) are measured exclusively on the training split; (iii) supervised baselines evaluated under the identical input serialisation and test protocol; and (iv) statistical validation of all headline comparisons through bootstrap confidence intervals and exact McNemar tests. We show that a medium-sized model, of 12–14B parameters, run locally can reach statistical parity with a supervised baseline trained on the full corpus [10].

Section II provides an overview of related works. Section III introduces our dataset, class engineering, candidate models and prompting methods. Section IV details the experimental setup and results. Section V concludes and lists future work.

II. State of the Art

Prior work falls into four groups: classical machine learning on physicochemical properties, neural classification of expert reviews, multimodal wine datasets, and prompt-based applications of large language models to wine-related text.

This research [11] conducts a leak-free study comparing five gradient-tree models on the UCI *Vinho Verde* dataset with eleven physicochemical attributes, reporting weighted F_1 scores of 0.693 (red) and 0.664 (white) for the best model, Gradient Boosting; restricting inputs to the five most predictive attributes costs at most three F_1 points, implying that physicochemical quality predictors are highly concentrated.

On the review-text side, [12] trains logistic regression on LDA, TF-IDF and Doc2Vec representations of $\sim$300K Wine Spectator reviews: Doc2Vec embeddings alone reach 83% binary accuracy, rising above 85% when critic scores, price and wine age are appended. Related work [13] analyses neural approaches on $\sim$140K reviews under 10-fold cross-validation: fine-tuned BERT achieves 89.12% (binary) and 81.57% (four-class) accuracy, narrowly surpassing BiLSTM and CNN baselines. On WineMag.com reviews, [14] trains hierarchical attention networks on $\sim$46K reviews, peaking at 84.71% validation accuracy with pretrained GloVe embeddings.

Tabular data are fed to LLMs by serialising table rows as natural-language prompts in [15]: evaluated on nine benchmark tabular datasets, TabLLM attains competitive or state-of-the-art few-shot classification accuracy with frozen instruction-tuned weights. This research [16] applies six open-weights and commercial LLMs to cross-cultural translation and substitution of wine reviews between English and Chinese (20973 WineEnthusiast and 4776 Weibo reviews), evaluated by BLEU (up to 18.7) and BERTScore (86.6–91.2). A recent survey in the Harvard Data Science Review [17] argues that prompt-based LLM evaluation of wine reviews remains comparatively understudied, with most prior work in computational enology focusing instead on lexicon-based sentiment analysis, document-level attention mechanisms or knowledge retrieval.

Demonstration retrieval for in-context learning is itself an active research area [18], where most work studies trainable retrievers and selection policies on balanced general-purpose NLP benchmarks. Our study is complementary on three axes absent from that literature and from the wine-NLP work above: frozen off-the-shelf retrieval applied to a severely imbalanced ordinal task in a specialised domain; a controlled ablation of class-balanced versus global exemplar selection under that imbalance; and protocol-matched supervised baselines with paired significance tests on consumer-grade hardware.

III. Methods

A. Dataset and Splits

Experiments are conducted on the *Wine Reviews Ordinal* dataset [19], an ordinal-classification adaptation of the WineMag.com expert review corpus originally compiled by Thoutt. The raw corpus of $N = 105154$ professional wine reviews contains four tabular attributes $x_{\text{country}}, x_{\text{region}}, x_{\text{grape}}, x_{\text{price}}$, a numerical quality score $r_i \in \{80, 81, \ldots, 100\}$ assigned by a sommelier on the WineEnthusiast point scale, and a free-text review body $\mathbf{t}_i \in \Sigma^*$. The dataset provides disjoint splits $\mathcal{D}_{\text{tr}}, \mathcal{D}_{\text{va}}, \mathcal{D}_{\text{te}}$ with sizes $71504/12619/21031$, which we consume directly.

Rows missing the target/review pair are discarded, as are reviews with fewer than ten alphanumeric characters. Missing categorical attributes are replaced by the constant token "Unknown"; missing prices by the literal string `unknown`. The grape-variety attribute is released as an integer code (0–29); we decode it to human-readable variety names via the categorical label mapping of the sibling (non-ordinal) release of the same corpus, verified by joining the two releases on the review text ($30000/30000$ matching codes). No further normalisation is applied since all classifiers below operate over plain-text input.

B. Label Engineering via Quality-Tier Binning

Wine scores are binned according to their associated review tier, one of six qualitative categories published in WineEnthusiast's editorial rubric: *acceptable* (80–82), *good* (83–86), *very good* (87–89), *excellent* (90–94), *superb* (95–97), and *classic* (98–100). We collapse this six-way ordinal scale into $K = 3$ groups determined exclusively by the magazine's thresholds:

$$y_i = \begin{cases} 0 \text{ (Low)} & \text{if } 80 \le r_i \le 82, \\ 1 \text{ (Medium)} & \text{if } 82 < r_i \le 90, \\ 2 \text{ (High)} & \text{if } 90 < r_i \le 100. \end{cases} \tag{1}$$

Fixed cut-off values $\{82, 90\}$ are determined a priori based on editorial criteria, rather than estimated from data.

Training-set frequencies are concentrated around the central score bands (86–88: 34.4%, 89–91: 30.3%), with low-density tails (80–82: 2.3%, 95–99: 2.0%), yielding the training prior $\Pr(y) \approx (0.023, 0.703, 0.274)$. We preserve this imbalance because it mirrors real-world skew in professionally-curated wine corpora; identical thresholds are applied to all splits.

C. Test Subsampling and Retrieval Pool

As LLM evaluation is bounded by inference latency, we assess each $(\text{model}, \text{strategy})$ pair on a stratified subsample $\mathcal{D}_{\text{te}}^{\text{eval}} \subset \mathcal{D}_{\text{te}}$ of nominal size $N_{\text{eval}} = 500$ (realised size 501 after per-class rounding), preserving the class distribution of $\mathcal{D}_{\text{te}}$ with $(11, 352, 138)$ examples per class:

$$\mathcal{D}_{\text{te}}^{\text{eval}} = \bigcup_{k=0}^{K-1} \text{Sample}\left(\mathcal{D}_{\text{te}}^{(k)}, \lceil N_{\text{eval}} \pi_k \rceil\right) \tag{2}$$

where $\mathcal{D}_{\text{te}}^{(k)} = \{(x, y) \in \mathcal{D}_{\text{te}} : y = k\}$ and π_k is the empirical class prior on $\mathcal{D}_{\text{tr}}$.

The retrieval pool used by the retrieval-augmented strategies is the *full* training split, $\mathcal{P} = \mathcal{D}_{\text{tr}}$. Pool size was selected on the validation split, never on test: on 300 stratified validation examples, retrieval from the full pool improved Mistral-Nemo-12B accuracy over a 3000-review stratified subsample by $+5.3$ percentage points (exact McNemar $p = 0.014$), as a $24\times$ denser pool yields substantially closer nearest neighbours. Increasing the exemplar count from $K_{\text{ret}} = 6$ to 12 *reduced* validation accuracy ($p = 0.029$), so $K_{\text{ret}} = 6$ is retained. The validation records are included in the project repository.

D. Frozen Sentence Embeddings for Retrieval

Each review $\mathbf{t}_i \in \mathcal{P} \cup \mathcal{D}_{\text{te}}^{\text{eval}}$ is encoded with the pretrained sentence transformer *BAAI/bge-large-en-v1.5*, fixing the model weights and ℓ_2-normalising the outputs:

$$\mathbf{e}_i = \frac{f_{\text{bge}}(\mathbf{t}_i)}{\|f_{\text{bge}}(\mathbf{t}_i)\|_2} \in \mathbb{R}^{1024}. \tag{3}$$

Cosine similarity between normalized embeddings simplifies to their dot product, enabling vectorized neighbourhood search at test time. We choose *bge-large-en-v1.5* for its strong performance on sentence-retrieval benchmarks; encoding the full 71.5k-review pool takes approximately ten minutes on the GPU, once per session.

E. Prompt Templates

All prompting strategies employ the same system prompt P^{sys}, which:

1) creates a sommelier persona;
2) lists the three quality classes with prototypical descriptor phrases distilled from the most discriminative features of a TF-IDF logistic regression fitted on $\mathcal{D}_{\text{tr}}$ (e.g. *thin*, *vegetal*, *harsh* for Low; *balanced*, *straightforward* for Medium; *complex*, *concentrated*, *long finish* for High);
3) states statistical priors measured exclusively on $\mathcal{D}_{\text{tr}}$: class base rates ($\approx 2/70/28\%$); a price ladder (under 10 USD wines almost never reach High; over 300 USD more than 95% are High); a review-length ladder (under 190 characters almost never High; 41% of High-class reviews exceed 300 characters); grape-variety and origin priors (Nebbiolo, Riesling, Austria and Germany lean High; Pinot Grigio and Rosé rarely reach High); and explicit guidance for the decisive Medium/High boundary;
4) instructs the model to output only a single digit in $\{0, 1, 2\}$, with no explanation, punctuation, whitespace, or additional text.

The input formatting function $\text{Wine}(\cdot)$ serialises a wine example as plain text containing the country, region, USD price, the decoded grape variety, the review character count, stated explicitly because autoregressive models cannot count characters reliably, and length is itself predictive of quality, and the full review text. Formatting is identical for each in-context example, followed by a suffix line `Class: k`.

F. Prompt Structures

1) Zero-shot: The user message contains only the serialised wine under classification, $P_i^{\text{zs}} = [P^{\text{sys}}; \text{Wine}(\mathbf{x}_i)]$, followed by an explicit single-digit instruction.

2) Few-shot with Random Exemplars: We uniformly sample ($K_{\text{fs}} = 4$ per class, fixed seed) 12 labelled exemplars from $\mathcal{P}$ per test instance and concatenate them under a uniform random permutation $\pi(\cdot)$ that prevents class-order priming: $P_i^{\text{fs}} = [P^{\text{sys}}; \pi(\mathcal{E}_i^{\text{rand}}); \text{Wine}(\mathbf{x}_i)]$.

3) Retrieval-Augmented Few-shot (Class-Balanced): Given test embedding $\mathbf{e}_i$, we build the exemplar set $\mathcal{E}_i^{\text{ret}}$ of $K_{\text{ret}} = 6$ nearest neighbours in $\mathcal{P}$ *balanced across classes*: for each class k we take the top-2 entries under cosine similarity,

$$\mathcal{E}_i^{\text{ret},k} = \arg\text{top-2}\Big(\big\{\mathbf{e}_i^\top \mathbf{e}_j : j \in \mathcal{P},\, y_j = k\big\}\Big), \tag{4}$$

concatenated in a shuffled order under a deterministic per-instance seed: $P_i^{\text{ret}} = [P^{\text{sys}}; \mathcal{E}_i^{\text{ret}}; \text{Wine}(\mathbf{x}_i)]$. Balancing is intended to prevent the severe class imbalance of $\mathcal{D}_{\text{tr}}$ ($\approx 70\%$ Medium) from flooding the exemplar set with Medium-class anchors; whether this constraint helps or instead sacrifices exemplar similarity is an empirical question, tested directly against the unbalanced variant below.

4) Retrieval-Augmented Few-shot (Global): As the ablation counterpart, the global variant selects the 6 highest-similarity exemplars regardless of class,

$$\mathcal{E}_i^{\text{glob}} = \arg\text{top-6}\Big(\big\{\mathbf{e}_i^\top \mathbf{e}_j : j \in \mathcal{P}\big\}\Big), \tag{5}$$

with an otherwise identical prompt. Comparing the two variants isolates the specific effect of class balancing from the effect of retrieval augmentation itself.

5) Self-consistency: Self-consistency is applied to the highest-accuracy (model, strategy) pair. The prompt is submitted $N_{\text{sc}} = 3$ times under stochastic decoding (temperature $\tau = 0.6$, top-$p = 0.95$, at most 4 output tokens), following the decoding parameters commonly adopted in the self-consistency literature [20]. Denoting by $\hat{y}_{i,n}$ the label predicted by the n-th sample, the final prediction is the mode:

$$\hat{y}_i^{\text{sc}} = \arg\max_{k \in \{0,1,2\}} \sum_{n=1}^{N_{\text{sc}}} \mathbf{1}[\hat{y}_{i,n} = k]. \tag{6}$$

G. Inference and Parsing

Inference is served using the Ollama Python client. Except for self-consistency, decoding is fully deterministic ($\tau = 0$, top-$p = 1.0$, at most 8 generated tokens). The serving context window is set to 4096 tokens, verified against a token-length audit of all constructed prompts (maximum $\approx$ 2.3k tokens, random few-shot), so that no prompt is silently truncated by the runtime's smaller default window. Response strings are mapped to labels by a regex matcher extracting the first standalone digit in $\{0, 1, 2\}$; unparseable responses would default to the median class and increment a failure counter, but in practice all 6012 deterministic responses parsed successfully.

H. Models and Supervised Baselines

We evaluate three open-weights instruction-tuned LLMs of 12–14B parameters, served at Q4_K_M ($\approx$ 4.7-bit) quantisation via Ollama: Qwen-2.5-14B-Instruct (Alibaba), Mistral-Nemo-12B-Instruct (Mistral AI), and Gemma-3-12B-Instruct (Google), three distinct lineages, so that behavioral differences reflect architectural and pretraining-corpus choices rather than a common source model. Each model is paired with each single-shot strategy, forming a 3×4 grid; self-consistency is run only for the single best cell.

As protocol-matched references, we additionally evaluate three supervised baselines on the identical serialisation $\text{Wine}(\cdot)$ and identical test subset: a majority-class predictor, and TF-IDF (word 1–2-grams, sublinear tf, up to 2×10^5 features) followed by logistic regression trained on the full $\mathcal{D}_{\text{tr}}$, in a default and a class-weight-balanced variant.

I. Metrics and Statistical Analysis

Seven metrics are computed over $\mathcal{D}_{\text{te}}^{\text{eval}}$. *Classification quality* is measured by accuracy and by the ordinal-aware mean absolute error,

$$\text{Acc} = \frac{1}{N_{\text{eval}}} \sum_i \mathbf{1}[\hat{y}_i = y_i] \tag{7}$$

$$\mathrm{MAE}_{\mathrm{ord}} = \frac{1}{N_{\mathrm{eval}}} \sum_i |y_i - \hat{y}_i| \tag{8}$$

where $\mathrm{MAE}_{\mathrm{ord}}$ penalises two-step confusions (Low↔High) twice as heavily as adjacent-class errors, together with the weighted and macro-averaged F_1 scores computed from per-class precision P_k and recall R_k; macro-F_1 weights all classes equally and is the most informative aggregate under strong imbalance. *Agreement and correlation* are captured by Cohen's chance-corrected kappa,

$$\kappa = \frac{p_o - p_e}{1 - p_e}, \quad p_o = \mathrm{Acc}, \quad p_e = \sum_k \frac{|\hat{\mathcal{C}}_k|}{N_{\mathrm{eval}}} \cdot \frac{|\mathcal{C}_k|}{N_{\mathrm{eval}}}, \tag{9}$$

by the multiclass Matthews correlation coefficient (MCC) computed from the confusion matrix, and by Spearman's rank correlation ρ_S between predicted and true labels. Lower is better for $\mathrm{MAE}_{\mathrm{ord}}$, higher for all remaining metrics.

Statistical analysis. For every configuration we report 95% bootstrap confidence intervals for accuracy and macro-F_1 (10^4 percentile resamples). Because all configurations predict on the same test items, pairwise comparisons use the exact two-sided McNemar test (binomial test on discordant prediction pairs), which has substantially higher power than comparing independent proportions at the same sample size.

IV. Results

A. Experimental Setup

Experiments are performed on a single laptop with an Intel Core 7 240H CPU (16 cores) and NVIDIA GeForce RTX 5050 Laptop GPU (8 GB VRAM). Ollama serves the LLMs; prompt construction, retrieval and metric computation run in a separate Python process. Random seed is 42 and batch size is 1 (request-serial inference). The implementation, including the experimental notebook, prompt templates, generated figures, processed results, the validation records for pool-size selection, and a complete worked example (raw input, fully assembled retrieval-augmented prompt, and raw model output), is available in the public project repository.[1]

B. Main Results

Table I reports all metrics, bootstrap confidence intervals and wall-clock times for the twelve (model, strategy) configurations, the self-consistency run, and the three supervised baselines. The highest-performing configuration is Gemma-3-12B with global retrieval at Acc $= 0.868$ $[0.838, 0.896]$, followed by Mistral-Nemo-12B with class-balanced retrieval at 0.862 $[0.832, 0.892]$. Both numerically surpass the strongest supervised baseline (TF-IDF + logistic regression, 0.854), although the paired difference is not statistically significant ($p = 0.435$): the prompted frozen LLMs reach statistical parity with a discriminative model trained on all 71504 labelled examples. Notably, the best LLM configuration also attains the highest macro-F_1 (0.744) of every evaluated system, including both TF-IDF variants. Accuracies across the twelve single-shot configurations span

[1] https://github.com/Pletea-Marinescu-Valentin/Wine-Quality

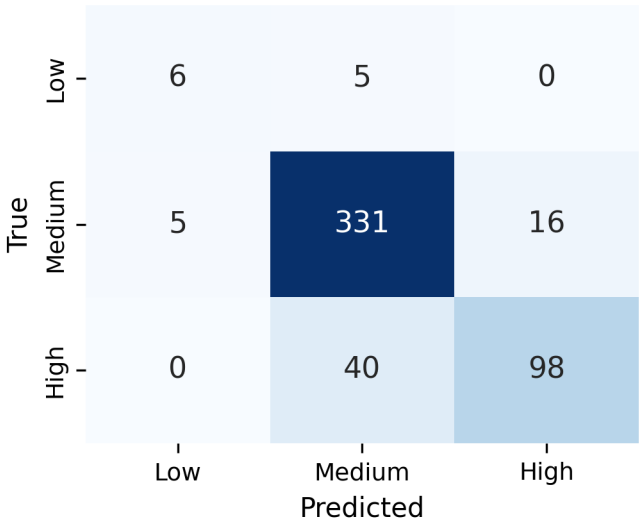

Fig. 1. Confusion matrix of the best configuration, Gemma-3-12B with global retrieval-augmented prompting (Acc $= 0.868$)

0.701–0.868. Every misclassification by every configuration is exactly one ordinal step (no Low↔High confusion occurs), hence $\mathrm{MAE}_{\mathrm{ord}} = 1 - \mathrm{Acc}$ throughout. Figure 1 presents the confusion matrix of the best configuration (confusion matrices for all sixteen configurations are available in the repository); the principal error mode is Medium–High confusion, reflecting the inherent ambiguity of the 90-versus-91-point boundary.

C. Effect of Prompting Strategy

Table II reports the key paired comparisons. Retrieval augmentation is the decisive factor: for every model family, the stronger retrieval variant exceeds zero-shot by $+5.6$ to $+7.8$ percentage points, and for Mistral-Nemo-12B the balanced-retrieval gain over zero-shot ($+8.0$ pp) is highly significant ($p < 0.001$). Random few-shot exemplars help marginally at best: they leave accuracy within ± 2.2 pp of zero-shot for Mistral-Nemo and Gemma-3 and *reduce* it by 8.6 pp for Qwen-2.5, while retrieved exemplars beat random exemplars significantly for all three models, semantically irrelevant exemplars provide noise, whereas retrieved neighbours serve as effective anchors for in-context calibration.

D. Class-Balanced versus Global Retrieval

The ablation yields a nuanced answer: the specific effect of class balancing is smaller than the effect of retrieval augmentation itself, and its sign is model-dependent. Global retrieval is significantly better for Gemma-3-12B ($+4.2$ pp, $p = 0.009$) and Qwen-2.5-14B ($+4.0$ pp, $p = 0.017$), while class balancing is better for Mistral-Nemo-12B ($+2.0$ pp, not significant). With the dense full-corpus pool, unconstrained top-K neighbours are already highly similar to the query, and forcing two exemplars per class trades similarity for coverage, a trade-off that pays off only for the model family that exploits

TABLE I

WINE QUALITY CLASSIFICATION RESULTS ON THE STRATIFIED TEST SUBSAMPLE ($N_{\text{EVAL}} = 501$), SORTED BY ACCURACY. ZS = ZERO-SHOT, FS = RANDOM FEW-SHOT, RET. = CLASS-BALANCED RETRIEVAL, RET.-G = GLOBAL RETRIEVAL, SC = SELF-CONSISTENCY, SUP. = SUPERVISED. TIME IS WALL-CLOCK MINUTES FOR ALL EVALUATION QUERIES (BASELINES: FIT + PREDICT).

Model	Strategy	Acc [95% CI]	F_1^w	F_1^m	κ	MCC	MAE_{ord}	ρ_S	Time
Gemma-3-12B	Ret.-G	**0.868** $[0.838, 0.896]$	**0.865**	**0.744**	**0.678**	**0.683**	**0.132**	0.711	23.8
Mistral-Nemo-12B	Ret.	0.862 $[0.832, 0.892]$	0.858	0.693	0.662	0.667	0.138	0.698	10.6
TF-IDF + LogReg	Sup.	0.854 $[0.822, 0.884]$	0.846	0.652	0.635	0.642	0.146	0.663	0.2
Gemma-3-12B (SC)	Ret.-G+SC	0.852 $[0.820, 0.882]$	0.848	0.715	0.636	0.643	0.148	0.677	45.2
TF-IDF + LogReg (bal.)	Sup.	0.848 $[0.816, 0.880]$	0.853	0.700	0.666	0.669	0.152	**0.713**	0.1
Qwen-2.5-14B	Ret.-G	0.846 $[0.814, 0.878]$	0.847	0.683	0.651	0.652	0.154	0.688	14.7
Mistral-Nemo-12B	Ret.-G	0.842 $[0.810, 0.874]$	0.831	0.619	0.591	0.610	0.158	0.638	10.7
Gemma-3-12B	Ret.	0.826 $[0.792, 0.860]$	0.838	0.667	0.614	0.618	0.174	0.695	23.8
Qwen-2.5-14B	Ret.	0.806 $[0.771, 0.840]$	0.817	0.654	0.587	0.594	0.194	0.657	14.5
Mistral-Nemo-12B	FS	0.804 $[0.769, 0.838]$	0.813	0.675	0.591	0.604	0.196	0.652	15.4
Gemma-3-12B	ZS	0.790 $[0.755, 0.826]$	0.801	0.618	0.536	0.538	0.210	0.614	16.9
Qwen-2.5-14B	ZS	0.786 $[0.751, 0.822]$	0.797	0.624	0.528	0.531	0.214	0.611	6.5
Gemma-3-12B	FS	0.786 $[0.751, 0.822]$	0.802	0.618	0.537	0.541	0.214	0.632	30.4
Mistral-Nemo-12B	ZS	0.782 $[0.747, 0.818]$	0.785	0.613	0.531	0.543	0.218	0.577	5.7
Majority class	Sup.	0.703 $[0.663, 0.743]$	0.580	0.275	0.000	0.000	0.297	0.000	< 0.1
Qwen-2.5-14B	FS	0.701 $[0.661, 0.741]$	0.738	0.556	0.410	0.425	0.299	0.592	22.4

TABLE II

EXACT MCNEMAR TESTS FOR KEY PAIRED COMPARISONS ($N = 501$ SHARED TEST ITEMS). POSITIVE Δ FAVOURS THE FIRST-LISTED CONFIGURATION.

Comparison	ΔAcc (pp)	p
Mistral: Ret. vs. ZS	$+8.0$	< 0.001
Qwen: Ret. vs. FS	$+10.6$	< 0.001
Mistral: Ret. vs. FS	$+5.8$	0.004
Gemma: Ret.-G vs. Ret.	$+4.2$	0.009
Qwen: Ret.-G vs. Ret.	$+4.0$	0.017
Gemma: Ret. vs. FS	$+4.0$	0.029
Mistral: Ret. vs. Ret.-G	$+2.0$	0.154
Gemma Ret.-G vs. TF-IDF + LogReg	$+1.4$	0.435
Gemma: Ret.-G vs. SC	$+1.6$	0.057

contrastive anchors most effectively. The strongest overall cell uses global retrieval (Gemma-3, 0.868), while class balancing gives Mistral-Nemo its best result (0.862).

E. Per-Class Behaviour and Minority-Class Diagnostic

For the strongest LLM configuration (Gemma-3-12B, global retrieval), per-class precision/recall are $0.55/0.55$ (Low), $0.88/0.94$ (Medium) and $0.86/0.71$ (High); for the balanced TF-IDF baseline, $0.33/0.55$, $0.92/0.86$ and $0.76/0.85$ respectively (full per-class tables for all sixteen configurations are provided in the repository). The dominant weakness of all systems is thus High-class recall (the Medium/High boundary), while the rare Low class (11 stratified instances) is measured too coarsely in the main subsample for reliable conclusions. We therefore ran a dedicated diagnostic on 100 additional Low-class reviews sampled from the full test split: the best LLM configuration recalls 0.48 $[0.39, 0.58]$ (Wilson 95% CI) of Low-class wines versus 0.77 $[0.68, 0.84]$ for the balanced TF-IDF baseline, a clear remaining gap. Notably, every missed Low-class instance is predicted Medium (the adjacent class) and none High, for both systems.

F. Effect of Model Family

No single family dominates: Gemma-3-12B achieves the best overall cell (with global retrieval) and is the most retrieval-responsive family ($+7.8$ pp over its zero-shot), Mistral-Nemo-12B leads under class-balanced retrieval, and Qwen-2.5-14B, despite the largest parameter count, never places first. Parameter count within the 12–14B range is thus not predictive of task performance: instruction-tuning data, domain coverage of the pretraining corpus, and decoding behaviour at temperature 0 evidently dominate.

G. Self-Consistency

Self-consistency applied to the best single-shot cell (Gemma-3-12B with global retrieval) yields 0.852, 1.6 pp *below* the deterministic baseline ($p = 0.057$). This null result is expected under our single-digit output protocol: with no chain-of-thought to marginalise over, sampling at $\tau = 0.6$ does not explore alternative reasoning paths but merely adds decoding noise around the deterministic argmax, flipping predictions predominantly on boundary instances. Practitioners should not expect chain-of-thought-free self-consistency to improve single-token classification.

H. Comparison with Prior Art

Our best configuration reaches 0.868 three-class accuracy with frozen 12B-parameter weights and no fine-tuning. Prior supervised results on cognate wine-review tasks (fine-tuned BERT at 89.12% binary accuracy on Wine Spectator data [13]; HAN at 84.71% validation accuracy on WineMag reviews [14]) are obtained on different corpora, class schemata, and evaluation protocols, so they serve as context rather than head-to-head comparisons. The protocol-matched comparison in this study is the TF-IDF baseline of Table I, with which the prompted LLMs achieve statistical parity while holding the better macro-F_1.

I. Caveats and Limitations

First, although the evaluation subsample was enlarged to $N_{\text{eval}} = 501$ and every headline comparison carries a confidence interval and a paired significance test, differences below ≈ 3 pp remain statistically unresolved; conclusions are drawn only from the significant comparisons of Table II. Second, class imbalance is severe (a majority-class predictor attains 0.703), so we accompany accuracy with macro-F_1, κ and MCC; the minority-class diagnostic nevertheless shows the supervised baseline remains substantially better at recalling Low-class wines (0.77 vs. 0.48). Third, all findings derive from a single corpus of English expert reviews; the pipeline is domain-agnostic, frozen embeddings, retrieval, and prompt priors re-derivable from any labelled training split transfer directly to other ordinal review-classification tasks, but empirical generalization remains to be demonstrated. Finally, retrieval-augmented prompting increases inference cost (10.6–23.8 minutes per 501 queries versus 4.4–16.9 for zero-shot), and the serving runtime's default context window (2048 tokens in Ollama) silently truncates longer prompts, so an explicit context-window configuration verified against a prompt-length audit is required for correct few-shot evaluation.

V. Conclusion

We presented a retrieval-augmented prompting pipeline for classifying wine quality from expert reviews into High, Medium, and Low categories. The best setup (Gemma-3-12B with global retrieval) achieved 0.868 accuracy $[0.838, 0.896]$ with the highest macro-F_1 (0.744) of all evaluated systems, retrieval augmentation delivered statistically significant gains over zero-shot and random few-shot prompting, and, without any gradient updates, the strongest prompted LLMs reached statistical parity with a TF-IDF logistic-regression baseline trained on the full 71.5k-review corpus. Class balancing in retrieval proved model-dependent, larger models within the 12–14B range did not provide stronger in-context learning, and chain-of-thought-free self-consistency consistently failed to improve deterministic decoding. The results support retrieval-augmented in-context learning over frozen open-weight LLMs as a practical alternative to gradient-based methods for low- to moderate-resource ordinal classification on commodity hardware. For future work we plan to: (i) move to more powerful, edge-oriented GPU infrastructure [21], [22] enabling larger instruction-tuned models (30–70B) and higher-precision inference beyond Q4_K_M; (ii) extend the label space to the full six-class WineEnthusiast scale under finer-grained ordinal distinctions; and (iii) fuse the reviews with physicochemical features (alcohol content, volatile acidity, residual sugar, pH) to test complementarity of qualitative and quantitative signals.

References

[1] M. B. Gačnik, A. Škraba, K. Pažek, and Č. Rozman, "Predicting wine quality under changing climate: An integrated approach combining machine learning, statistical analysis, and systems thinking," *Beverages*, vol. 11, no. 4, p. 116, 2025.

[2] Y. Gupta and A. Saraswat, "Wine quality prediction based on machine learning techniques," in *Flexible Electronics for Electric Vehicles: Select Proceedings of FlexEV—2021*. Springer, 2022, pp. 623–627.

[3] P. Manikanta, D. Jayaprakash, D. Sharma, and R. Bathla, "Wine quality prediction using machine learning techniques," in *2024 2nd International Conference on Advances in Computation, Communication and Information Technology (ICAICCIT)*, 2024, pp. 1015–1018.

[4] A. A. S. Pradhana, K. S. Batubulan, and I. N. D. Kotama, "Predicting wine quality based on features using naive bayes classifier," *Jurnal Sistem Informasi dan Komputer Terapan Indonesia (JSIKTI)*, vol. 7, no. 1, pp. 275–284, 2024.

[5] H. Arshad, "The wine quality prediction using machine learning," *J Innovative Computing and Emerging Technologies*, vol. 4, no. 2, 2024.

[6] I. Ilieva, M. Terziyska, and T. Dimitrova, "From words to ratings: Machine learning and nlp for wine reviews," *Beverages*, vol. 11, no. 3, p. 80, 2025.

[7] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," *Advances in neural information processing systems*, vol. 33, pp. 9459–9474, 2020.

[8] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[9] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, 2019, pp. 3982–3992.

[10] A. Edwards and J. Camacho-Collados, "Language models for text classification: Is in-context learning enough?" in *2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 10 058–10 072.

[11] Z. Chen, "Wine quality prediction with ensemble trees: A unified, leak-free comparative study," in *Proceedings of the 2nd International Conference on Image Processing, Machine Learning, and Pattern Recognition*, 2025, pp. 162–170.

[12] C. Yang, J. Barth, D. Katumullage, and J. Cao, "Wine review descriptors as quality predictors: Evidence from language processing techniques," *Journal of Wine Economics*, vol. 17, no. 1, pp. 64–80, 2022.

[13] D. Katumullage, C. Yang, J. Barth, and J. Cao, "Using neural network models for wine review classification," *Journal of Wine Economics*, vol. 17, no. 1, pp. 27–41, 2022.

[14] V. Diaconita, A. Belciu, A. M. I. Corbea, and I. Simonca, "Decoding wine narratives with hierarchical attention: Classification, visual prompts, and emerging e-commerce possibilities," *Journal of Theoretical and Applied Electronic Commerce Research*, vol. 20, no. 3, p. 212, 2025.

[15] S. Hegselmann, A. Buendia, H. Lang, M. Agrawal, X. Jiang, and D. Sontag, "Tabllm: Few-shot classification of tabular data with large language models," in *International conference on artificial intelligence and statistics*. PMLR, 2023, pp. 5549–5581.

[16] C. Zou, X. Wen, T. Hu, Q. J. Wang, and D. Hershcovich, "Do llms understand wine descriptors across cultures? a benchmark for cultural adaptations of wine reviews," *arXiv preprint arXiv:2509.12961*, 2025.

[17] J. Cao, "Ai: A new and impactful player in the quality evaluation of wine," *Harvard Data Science Review*, vol. 7, no. 1, 2025.

[18] M. Luo, X. Xu, Y. Liu, P. Pasupat, and M. Kazemi, "In-context learning with retrieved demonstrations for language models: A survey," *arXiv preprint arXiv:2401.11624*, 2024.

[19] J. Burton, "Wine reviews ordinal dataset," 2023. [Online]. Available: https://huggingface.co/datasets/james-burton/wine_reviews_ordinal

[20] X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery, and D. Zhou, "Self-consistency improves chain of thought reasoning in language models," *arXiv preprint arXiv:2203.11171*, 2022.

[21] M.-D. Pavel, G. Stamatescu, M. Chodnicki, and C. G. Amza, "Llm-enhanced control of a mobile robotic platform for smart industry," *Applied Sciences*, vol. 16, no. 4, 2026. [Online]. Available: https://www.mdpi.com/2076-3417/16/4/1680

[22] M.-D. Pavel, M. De Marchi, and G. Stamatescu, "Edge-cloud offloading for real-time perception of resource-constrained amr in iiot environments," in *2026 22th International Conference on Distributed Computing in Smart Systems and the Internet of Things (DCOSS-IoT)*.