A Controlled Benchmark of Open-Source and Proprietary LLMs for Few-Shot Microbiome IBD Classification

Nouhaila En Najih[1], Soufiane Hamida[2], Ahmed Moussa[3]
Systems and Data Engineering Team-National School of Applied Sciences,
Abdelmalek Essaadi University, Tangier, Morocco[1,3]
2IACS Laboratory, ENSET Mohammedia, Hassan II University of Casablanca, Mohammedia, Morocco[2]
GENIUS Laboratory, SupMTI of Rabat, Rabat, Morocco[2]

***Abstract*—Phenotyping inflammatory bowel disease (IBD) from gut microbiome profiles remains challenging due to 93% genus-level zero-inflation, skewed amplicon count distributions, and the practical cost of assembling labelled cohorts. Few-shot in-context learning (ICL) with large language models (LLMs) sidesteps the annotation bottleneck, yet existing benchmarks test only proprietary APIs—a deployment model incompatible with the data governance constraints of most clinical sites. We benchmark six frontier LLMs under identical few-shot conditions—three proprietary (GPT-4o, Claude claude-opus-4-5, Gemini 2.5 Flash Lite) and three open-source (Mistral 7B, LLaMA-3 8B, DeepSeek-R1-Distill-Qwen 1.5B)—evaluated against supervised baselines (Random Forest, XGBoost, LightGBM, soft-voting Ensemble) on 16S rRNA amplicon data ($n = 1{,}316$; holdout $n = 30$). All six LLMs received identical prompts, the same log-normalised top-20 features, and the same random shot selection strategy with a fixed seed, ensuring that observed differences are attributable to model capacity rather than experimental conditions. The supervised Ensemble led holdout performance (Macro-F1: 0.7948; AUC: 0.7725). Among LLMs, Mistral 7B achieved the highest Macro-F1 (0.5417), surpassing all three proprietary models without any parameter update. The mean performance gap between open-source (0.4711) and proprietary (0.5101) groups was only 0.039 Macro-F1 points—too narrow to justify exclusive reliance on commercial APIs in privacy-sensitive deployments. Within-open-source variance (0.130 Macro-F1 points) substantially exceeded this inter-family gap, indicating that model selection within the open-source ecosystem is the more consequential practical decision. These results suggest that open-source LLMs running on local hardware are a workable option when labeled data is limited and routing patient-derived data to external APIs is not permitted.**

***Keywords*—Inflammatory bowel disease; gut microbiome; few-shot learning; large language models; open-source LLMs; proprietary LLMs; in-context learning; 16S rRNA; log-normalisation; zero-inflation; GPT-4o; Claude; Gemini; LLaMA; Mistral; DeepSeek**

I. Introduction

Inflammatory bowel disease (IBD) affects an estimated 6.8 million people globally as of 2017, with incidence still rising in newly industrialising countries [2]. Its two main subtypes—Crohn's disease (CD) and ulcerative colitis (UC)—differ substantially in anatomical distribution, histological pattern, and treatment trajectory, but symptom overlap makes clinical differentiation unreliable without additional investigation. 16S rRNA amplicon sequencing captures gut compositional shifts non-invasively, and dysbiosis patterns linked to IBD activity have been reproduced across independent cohorts [3].

Supervised classifiers perform well when labelled data are plentiful [4]. The bottleneck is annotation cost: building a quality-controlled, labelled microbiome cohort for each clinical site or sequencing protocol is expensive, slow, and often not feasible in resource-limited settings. Few-shot ICL with LLMs sidesteps this requirement—a model pre-trained on biomedical literature can be adapted to a new classification task from a handful of labelled examples in the prompt, without any weight update [5]. What the field has not examined is whether this remains true when proprietary APIs are swapped for locally-deployed open-source models.

That gap has practical consequences. GDPR in Europe, HIPAA in the United States, and equivalent frameworks elsewhere routinely prohibit routing patient-derived biological profiles to external cloud services. For institutions operating under these constraints, proprietary APIs are simply unavailable, and the operational question reduces to: can an open-source model running on local hardware match the diagnostic utility of a commercial API? No existing study provides a controlled answer on sparse, zero-inflated biomedical tabular data.

This study addresses that gap directly by asking two explicit research questions: (RQ1) Is there a meaningful performance difference between proprietary and open-source LLMs on zero-inflated 16S rRNA classification data under identical few-shot conditions? (RQ2) Does within-family variance among open-source models exceed the cross-family gap, and if so, what architectural factors explain it? We evaluate six LLMs—three proprietary and three open-source—on 1,316 samples drawn from two publicly available IBD cohorts, using the same log-normalised microbiome features, the same random shot selection strategy, and the same supervised baselines as reference.

The remainder of this study is organised as follows. Section II reviews related work on supervised microbiome classification, clinical LLMs, and tabular few-shot learning. Section III details data, preprocessing, prompt design, and model configurations. Section IV presents results. Section V discusses implications and limitations. Section VI concludes.

This study makes three contributions: 1) the first head-to-head proprietary vs. open-source benchmark conducted

under rigorously identical prompt, encoding, and shot-selection conditions on zero-inflated 16S rRNA classification data; 2) quantification of the performance cost of local deployment; and 3) evidence that the within-open-source model spread (0.130 Macro-F1 points) exceeds the across-family gap (0.039 points), shifting practical guidance toward model selection within the open-source ecosystem rather than toward proprietary APIs.

II. Related Work

A. Supervised Classification of Gut Microbiome Profiles

The canonical preprocessing pipeline for 16S rRNA disease phenotyping combines total sum scaling (TSS) normalisation, log-transformation, and tree-based classifiers [4]. Pasolli et al. [4] surveyed machine learning strategies across 28 case-control metagenomic datasets and found Random Forest to be consistently competitive, yet highlighted cross-cohort generalisation as the field's central weakness: a meta-analysis found that a substantial fraction of differentially abundant genera reflect non-specific dysbiosis rather than disease-specific biology [6]. This batch sensitivity directly motivates annotation-free approaches that do not assume training and test cohorts share the same technical characteristics. More broadly, ensemble tree-based classifiers (Random Forest, XGBoost, LightGBM) have been consistently among the strongest performers on biomedical tabular tasks across diverse clinical domains, including diabetes prediction [29] and cardiovascular disease diagnosis under class imbalance [30], which informs our selection of supervised baselines in this study.

B. LLMs in Clinical Contexts

Frontier LLMs now perform at or above physician level on standardised clinical licensing examinations [7]. However, strong benchmark scores do not automatically transfer to structured biomedical data tasks. Van Veen et al. [8] demonstrated that adaptation is critical: models fine-tuned on clinical notes substantially outperformed general-purpose systems on summarisation tasks, suggesting that biomedical pre-training alone is insufficient for specialised domains. Similarly, Singhal et al. [9] showed that while LLMs encode broad clinical knowledge, performance on structured reasoning tasks involving numerical biomarkers remains uneven. Recent work has begun to apply LLMs directly to disease classification from structured clinical and demographic features: Almalki et al. [31] reported high diagnostic accuracy for Alzheimer's disease using GPT-3.5 on ADNI cognitive and biomarker data, demonstrating that LLMs can extract diagnostic signal from heterogeneous tabular biomedical inputs without explicit feature engineering. The few-shot ICL mechanism underlying these results [5] is well-studied for natural language tasks but poorly understood for high-sparsity biomedical tabular data—a gap this study targets directly.

C. LLMs on Tabular and Structured Data

TabLLM [10] established that GPT-3 can classify clinical tabular records in a few-shot regime when inputs are serialised as natural language sentences, reporting sparsity levels up to approximately 40% missing values in their benchmark. LIFT [11] showed that fine-tuning language models on numerical feature representations—without architectural changes—yields competitive performance across classification and regression benchmarks. Triantafillou et al. [12] separately documented that input encoding choices matter at least as much as algorithm selection in low-data regimes.

Critically, none of these studies operated at the extreme sparsity imposed by amplicon count data (93% zeros in our setting, well beyond the regime explored by TabLLM), and none compared proprietary and open-source deployments under identical conditions. Recent work has begun to compare open-source and proprietary LLMs in other domains—including clinical text and code generation—and consistently finds that the performance gap is smaller than assumed, with open-source models competitive on domain-specific tasks when model selection is careful [8], [9]. Our study extends this line of evidence to the specific and more demanding regime of zero-inflated compositional tabular data.

D. Recent LLM Applications in Microbiome Research

Recent work has started applying LLMs to microbiome tasks directly, though with limitations that matter for our setting.

Liu et al. [23] reported 73.91% accuracy on a single IBD cohort using a full labelled training set. With 11 in-context examples and no parameter update, Mistral 7B reaches 56.67% in our benchmark, while the supervised Ensemble reaches 83.33%. The difference between Mistral 7B and Liu et al. is primarily a function of how much labelled data each approach consumes, not of model capacity. A stricter comparison is precluded by differences in cohort composition and evaluation protocol across the two studies. Mu et al. [24] found classical ML consistently competitive with foundation models across 83 microbiome cohorts, consistent with our observation that the supervised Ensemble outperforms all six LLMs by a substantial margin.

Crucially, their framework does not compare proprietary vs. open-source deployments and does not quantify the annotation cost. Cao et al. [28] introduced Chat2GM, a hybrid ML–LLM system for gut microbiome analysis, but evaluated it exclusively on a mouse obesity dataset, limiting its clinical relevance. Park et al. [27] fine-tuned a domain-specific LLM (METABOLISM) on 160k biomedical abstracts for microbiome–liver interaction reasoning, but the system takes literature as input rather than patient-level microbiome profiles and does not classify disease from biological data. In the broader medical-AI literature, LLMs have also been integrated as components of structured diagnostic pipelines: Li and Sun [32] embedded a quantised LLaMA-2-7B as a multi-expert feedback module within a hierarchical reinforcement-learning framework for symptom-driven disease diagnosis, illustrating the expanding range of LLM-augmented medical decision-support systems.

At the benchmark level, Mu et al. [24] compared classical ML, GPT-derived embeddings, and foundation models (TabPFN, MGM) across 83 cohorts using 16S and shotgun data. Classical ML remained competitive throughout, and LLM-based embeddings did not consistently beat established methods—consistent with what we observe here. That said, Mu et al. do not test few-shot ICL on raw microbiome features, and do not put proprietary and open-source LLMs side-by-side under identical conditions. Reviews by Xing et al. [25] and Yan et

al. [26] point to growing interest in LLMs for microbiome sequence analysis and clinical support, and both flag the same two weaknesses: no standardised benchmarks and thin external validation—which is precisely what this study targets.

Table I situates this study among the closest prior work across four axes: task type, sparsity regime, few-shot paradigm, and open-source vs. proprietary comparison.

TABLE I. POSITIONING RELATIVE TO PRIOR WORK

Study	Task	Sparsity	Few-shot	OS vs Prop
TabLLM [10]	Clinical tab.	$\leq$40%	Yes	No
LIFT [11]	General tab.	Low	No	No
Singhal et al. [9]	Med. QA	N/A	Yes	No
Liu et al. [23]	Microbiome IBD	High	Yes	No
Mu et al. [24]	Microbiome	High	No	No
This work	Microbiome	**93%**	**Yes**	**Yes**

III. MATERIALS AND METHODS

A. Data and Preprocessing

The analysis draws on publicly available stool microbiome data from the European Nucleotide Archive (ENA; accession numbers PRJEB13679, PRJEB33711), totalling 1,316 samples across two retained cohorts (Table II). Raw FASTQ files were processed with DADA2 (v1.24) [1] in R (v4.5.1) following a unified single-end amplicon pipeline applied jointly to all samples: quality filtering (`truncLen=140`, `maxN=0`, `maxEE=2`, `truncQ=2`, `rm.phix=TRUE`), error-rate learning on a random subset of 25 samples (`set.seed(1)`) to reduce memory requirements, sample-wise denoising, chimera removal by consensus (`removeBimeraDenovo`), and taxonomic assignment against the SILVA database (v138.1). Genus-level aggregation ensured cross-cohort taxonomic harmonisation. No rarefaction was applied; library size variation was addressed downstream by TSS normalisation.

TABLE II. DATASET COMPOSITION

Batch	N	Control	CD	UC
Dataset_1	1,286	336 (26.1%)	731 (56.8%)	219 (17.1%)
Dataset_30	30	9 (30.0%)	10 (33.3%)	11 (36.7%)
Total	**1,316**	**345 (26.2%)**	**741 (56.3%)**	**230 (17.5%)**

At 93.4% zero-inflation—675 of the 799 genera individually exceed 90% zeros—prompts listing all genera are noise-dominated before reaching a single informative entry. Preprocessing follows [4]: TSS normalisation followed by log-CPM transformation $\tilde{x}_{ij} = \log_e\left(1 + x_{ij} \times 10^6\right)$. The top-20 genera by Random Forest mean-decrease-in-impurity, computed exclusively on Dataset_1 to avoid any leakage to the test set, serve as fixed features for both the LLM prompts and the supervised baselines, providing a shared biological prior across paradigms. The strongest discriminant is Roseburia (Kruskal–Wallis $H = 144.6$, $p < 10^{-30}$), followed by Agathobacter ($H = 105.5$) and Faecalibacterium ($H = 83.0$) [13], [14]. Dataset_1 ($n = 1{,}286$) served as the training and few-shot source; Dataset_30 ($n = 30$) served as the independent holdout.

B. Methodology

We evaluated four encoding strategies in a preliminary experiment on a held-aside subset of Dataset_1: P1 (raw integer counts), P2 (relative abundances), P3 (log-CPM values restricted to the top-10 non-zero features, normalised to sum to 1 within the prompt), and P4 (clinical narrative description). Macro-F1 across shot counts was highest and most stable for P3 (peak 0.530 at $k = 5$), followed by P2 (0.430 at $k = 3$), P1 (0.460 at $k = 5$), and P4 which performed inconsistently across models. We therefore adopt P3 as the fixed encoding strategy across all six LLMs, ensuring that observed performance differences are attributable to model capacity rather than representation choice.

The P3 encoding was selected for three reasons: log-CPM transformation reduces the dynamic range of amplicon counts and limits the token budget consumed by high-abundance genera; restricting to the top-10 non-zero features per sample removes uninformative zero-inflated entries from the prompt; and within-prompt normalisation to sum to 1 ensures that relative abundances are directly comparable across samples regardless of sequencing depth.

Each prompt embeds a system message encoding nine IBD dysbiosis patterns and directional decision rules (Roseburia $> 0.20 \rightarrow$ Control; Faecalibacterium < 0.05 AND Roseburia $< 0.05 \rightarrow$ IBD; uncertain cases default to IBD), followed by k labelled examples and the target sample serialised as a normalised JSON of non-zero relative abundances among the top-20 genera. An illustrative example prompt is provided in Fig. 1. The mean prompt length was approximately 1,800 tokens per inference call (including system message and 11 few-shot demonstrations), well within the context window of all six models.

```
[SYSTEM]
  - Role: biomedical AI classifier
  - Class prior: 70% IBD prevalence
  - 9 dysbiosis patterns with directional rules
    (genus thresholds -> Healthy / IBD)
  - Default: IBD if uncertain
  - Output format: JSON array only

[USER - demonstration 1]
  {"sample_id": "...",
   "features": {top-10 genera: rel. abund.}}
[ASSISTANT - demonstration 1]
  [{"sample_id":"...",
    "predicted_label":0,
    "confidence":...}]
  ... (k demonstrations total) ...

[USER - target]
  {"sample_id": "...",
   "features": {top-10 genera: rel. abund.}}
```

Fig. 1. Few-shot prompt structure (P3 encoding): system block with class priors and dysbiosis rules, $k = 11$ labelled demonstrations, then the target sample. Genus values are log-CPM normalised relative abundances of the top-10 non-zero features.

Examples were drawn by stratified random sampling from Dataset_1, with a fixed seed (`random_state=42`) for reproducibility. Shot composition was set asymmetrically at 3 Healthy + 8 IBD to match Dataset_30 class prevalence (70% IBD) and counteract the Healthy-prediction bias seen in preliminary trials, yielding 11 demonstrations per inference call. The fixed seed guarantees that all six models received the

same demonstration set, so observed differences in classification performance reflect model capacity rather than example selection.

The 3 Healthy / 8 IBD split was set to match the 70% IBD prevalence of Dataset_30 and correct for the healthy-prediction bias we observed when using balanced shot selection in preliminary trials. The cutoff at 20 genera was chosen empirically because the cumulative impurity importance plateaued after the top 20 features and no Macro-F1 improvement was observed when extending the feature set to 30 or 50 genera. The top-20 features come from Random Forest mean-decrease-in-impurity computed on Dataset_1 only, so no information from the holdout influenced which genera were selected [4].

C. Models

We evaluated six LLMs across two deployment families. The proprietary group comprised GPT-4o [15], Claude claude-opus-4-5 [16], and Gemini 2.5 Flash Lite [17], each queried through its vendor API. The open-source group comprised Mistral 7B Instruct v0.3 [18], LLaMA-3 8B Instruct [19], and DeepSeek-R1-Distill-Qwen 1.5B [20], run locally on an NVIDIA T4 GPU (15 GB VRAM, Google Colab) after loading with 4-bit NF4 quantisation to satisfy memory constraints. Exact model identifiers are reported in Table III.

All models received identical prompts at temperature $=$ 0.1. Proprietary models were called with `max_tokens=512`; local models used `max_new_tokens=256`. A multi-stage JSON parsing routine with up to five retries on malformed output handled cases where the response did not conform to the required array format; retries used exponential back-off (base wait $= 10 \times 2^{\text{attempt}}$ seconds for rate-limit errors, $5 \times 2^{\text{attempt}}$ for other errors). DeepSeek-R1 required an explicit output-format constraint to prevent chain-of-thought reasoning from exhausting the token budget before producing the JSON label; the decision rules were preserved identically across all six models.

D. Supervised Baselines

Random Forest ($n_{\text{est}} = 500$, $\max_{\text{depth}} = 10$, `random_state` $= 42$), Extra Trees ($n_{\text{est}} = 500$, $\max_{\text{depth}} = 10$, `random_state` $= 42$), XGBoost ($n_{\text{est}} = 200$, $\max_{\text{depth}} = 6$, $lr = 0.05$, `random_state` $= 42$), LightGBM ($n_{\text{est}} = 200$, $lr = 0.05$, `random_state` $= 42$), and a soft-voting Ensemble (RF+XGB+LGBM) were trained on Dataset_1 and evaluated on Dataset_30. Hyperparameters followed established practice in microbiome classification [4] and were held constant across all experiments. Library versions were `scikit-learn 1.6.1`, `xgboost 3.2.0`, and `lightgbm 4.6.0`. Class imbalance was handled via `class_weight="balanced"` for RF, Extra Trees, and LightGBM, and `scale_pos_weight=0.35` for XGBoost. We assessed internal stability on Dataset_1 using a class-stratified five-fold cross-validation scheme, with fold assignments shuffled before splitting and a fixed seed (`random_state` $= 42$) for reproducibility.

Macro-F1 is the primary metric given the asymmetric clinical cost of missed IBD diagnoses [21]. AUC is reported for supervised models only—LLM confidence scores are not calibrated probability estimates.

IV. Results

A. Supervised Baseline Performance

Five-fold stratified cross-validation on Dataset_1 confirmed stable rankings across folds (Table IV). The soft-voting Ensemble recorded the highest internal performance (AUC: 0.8499 ± 0.0269; Macro-F1: 0.7476 ± 0.0242), closely followed by LightGBM (AUC: 0.8483 ± 0.0249) and XGBoost (AUC: 0.8418 ± 0.0344). On the holdout, the Ensemble led all classifiers (Accuracy: 0.8333; Macro-F1: 0.7948; AUC: 0.7725). The cross-validation to holdout AUC drop was sharpest for Random Forest ($\Delta\text{AUC} = 0.083$), reflecting its greater sensitivity to batch-induced variance in the absence of explicit correction. Ensemble and boosting models were more robust to this shift, with ΔAUC below 0.09 for all three. The AUC comparison across folds and holdout is visualised in Fig. 2.

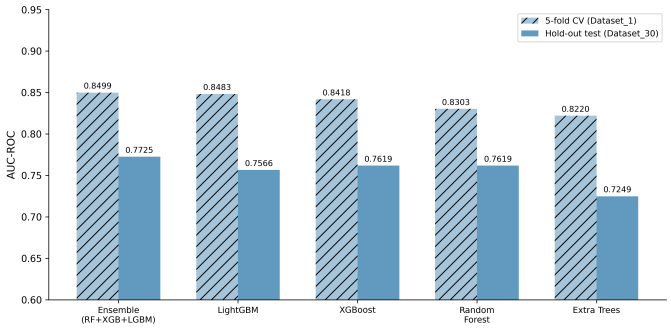

Fig. 2. AUC-ROC: 5-fold CV (Dataset_1) vs. hold-out (Dataset_30). The gap is sharpest for Random Forest, reflecting batch-induced variance.

B. Proprietary vs. Open-Source LLM Comparison

All six LLMs cleared the random baseline for a binary 70/30 classification task (Macro-F1 ≈ 0.41) without any parameter update, confirming that biomedical pre-training encodes transferable microbiome–IBD associations. Among proprietary models, Gemini 2.5 Flash Lite achieved the highest point-estimate Macro-F1 (0.5333; Accuracy: 0.5333), followed by GPT-4o (Macro-F1: 0.4994; Accuracy: 0.5667) and Claude opus-4-5 (Macro-F1: 0.4976; Accuracy: 0.5333), yielding a group mean of 0.5101. Bootstrap 95% confidence intervals overlap substantially across all three proprietary models (GPT-4o: [0.33, 0.67]; Claude: [0.32, 0.67]; Gemini: [0.27, 0.63]), indicating that no single proprietary model is distinguishably superior at $n = 30$. Among open-source models, Mistral 7B achieved the highest score (Macro-F1: 0.5417; Accuracy: 0.5667; 95% CI: [0.37, 0.72]), surpassing all three proprietary models. DeepSeek-R1-Distill-Qwen 1.5B recorded Macro-F1: 0.4118 (Accuracy: 0.4643; 95% CI: [0.32, 0.69]) and LLaMA-3 8B recorded Macro-F1: 0.4599 (Accuracy: 0.6333; 95% CI: [0.32, 0.64]), yielding a group mean of 0.4711. The inter-group gap was 0.039 Macro-F1 points, with all pairwise CIs overlapping, precluding any strong conclusion about family-level superiority. Against the closest published result (Liu et al. [23], 73.91% accuracy with a full training cohort), the best performing LLM in our benchmark (Mistral 7B, 56.67% accuracy from 11 demonstrations) confirms that few-shot ICL incurs a meaningful accuracy cost compared to fully supervised frameworks, while avoiding the need for large labelled training datasets.

TABLE III. MODEL IDENTIFIERS

Model	Family	Deployment	Identifier
GPT-4o	Proprietary	API	`gpt-4o`
Claude claude-opus-4-5	Proprietary	API	`claude-opus-4-5`
Gemini 2.5 Flash Lite	Proprietary	API	`gemini-2.5-flash-lite`
Mistral 7B Instruct v0.3	Open-source	Local	`mistralai/Mistral-7B-Instruct-v0.3`
LLaMA-3 8B Instruct	Open-source	Local	`unsloth/llama-3-8b-Instruct`
DeepSeek-R1-Distill 1.5B	Open-source	Local	`deepseek-ai/DeepSeek-R1-Distill-Qwen-1.5B`

TABLE IV. 5-FOLD CROSS-VALIDATION ON DATASET_1 ($n = 1{,}286$); MEAN $\pm$ STD.

Model	AUC (mean $\pm$ std)	Macro-F1 (mean $\pm$ std)
Ensemble (RF+XGB+LGBM)	0.8499 ± 0.0269	0.7476 ± 0.0242
LightGBM	0.8483 ± 0.0249	0.7394 ± 0.0189
XGBoost	0.8418 ± 0.0344	0.7450 ± 0.0171
Random Forest	0.8303 ± 0.0267	0.6961 ± 0.0187
Extra Trees	0.8220 ± 0.0322	0.7266 ± 0.0378

Full results are reported in Table V, with visual comparisons in Fig. 3 and Fig. 4. Accuracy and Macro-F1 per LLM are compared in Fig. 5, illustrating the divergence between the two metrics under class imbalance.

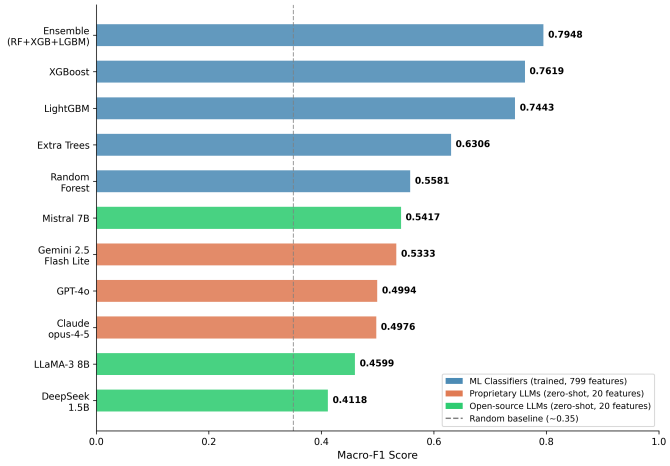

Fig. 3. Macro-F1 on hold-out (Dataset_30, $n = 30$). Dotted line: random baseline ($\approx$0.41). Blue: ML classifiers (799 features). Orange: proprietary LLMs. Green: open-source LLMs (few-shot, 20 features).

Two qualitatively different regimes are visible in this figure. The supervised classifiers spread across nearly a quarter of the Macro-F1 scale (0.56–0.79), with the soft-voting Ensemble at the upper edge and Random Forest at the lower one—a range that reflects genuine algorithmic differentiation when 1,286 labelled samples drive the learning. The six LLMs, in contrast, occupy a tight band of barely 0.13 Macro-F1 points (0.41–0.54): each clears the random baseline, yet none approaches the supervised range. This compression is itself informative. With only 11 in-context demonstrations and 20 features available to the LLMs, model capacity appears to become a secondary determinant of performance, with prompt encoding and feature selection absorbing most of the variation that algorithm choice would otherwise account for. We examine this hypothesis directly in the head-to-head LLM comparison shown in Fig. 4.

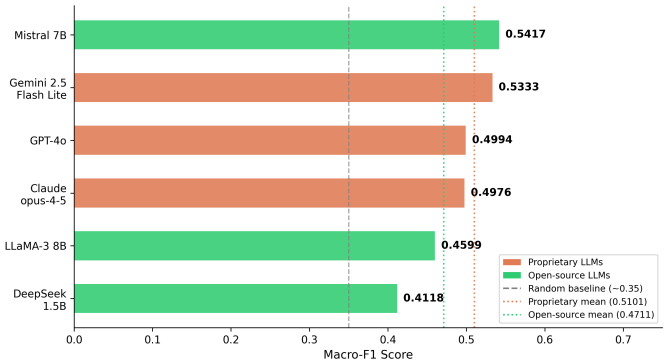

Fig. 4. Head-to-head comparison of the six LLMs. Dotted lines: proprietary mean (0.5101) and open-source mean (0.4711). Mistral 7B (open-source) outperforms all proprietary models.

What stands out in this figure is a finding that runs against the dominant intuition in the field: the most performant model on this task is also one of the smallest and the most freely accessible. Mistral 7B—7 billion parameters, permissive licence, deployed here at 4-bit NF4 precision on a single 15 GB GPU—is the only model whose point estimate crosses above the proprietary group mean. At the opposite end, DeepSeek-R1-Distill 1.5B, the smallest model evaluated, falls below both group means and pulls the open-source average down despite Mistral's lead. GPT-4o, Claude opus-4-5, Gemini 2.5 Flash Lite, and LLaMA-3 8B sit in a tight cluster around 0.50 Macro-F1 between these two extremes, a regime in which provenance—commercial or open—loses much of its predictive value for performance. The 0.039-point inter-family gap that emerges from this configuration is, for practical purposes, indistinguishable from the variance one would expect at $n = 30$ with overlapping bootstrap intervals. The accuracy/F1 divergence shown in Fig. 5 explains why this aggregate view can be misleading when accuracy alone is reported.

The LLaMA-3 8B case is a cautionary illustration of why aggregate metrics mislead in imbalanced settings. With 70% of holdout samples labelled IBD, a classifier that defaults to the majority class whenever uncertain will mechanically inflate its accuracy without ever learning to discriminate the minority class. LLaMA-3's 0.6333 accuracy paired with 0.4599 Macro-F1 is precisely the signature of this behaviour: the model captures the dominant phenotype well enough to look

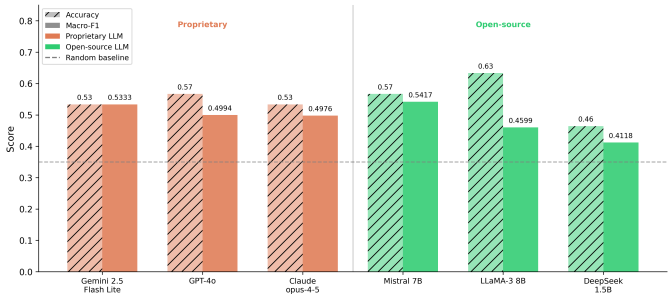

Fig. 5. Accuracy and Macro-F1 for the six LLMs by family. The two metrics diverge under class imbalance (70% IBD); Macro-F1 is the more reliable indicator.

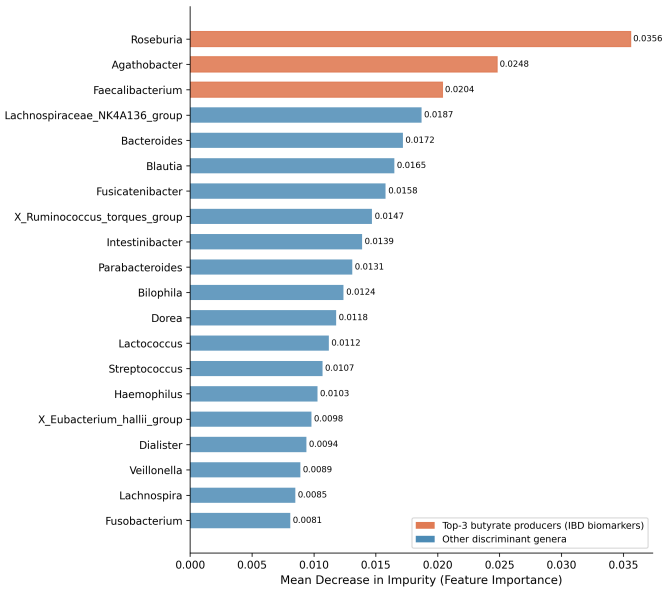

Fig. 6. Top-20 discriminant genera (Random Forest mean-decrease-in-impurity, Dataset_1, $n = 1{,}286$). Orange: top-3 butyrate producers (Roseburia, Agathobacter, Faecalibacterium) whose IBD-associated depletion is independently replicated [13], [14]. All 20 genera are shared with the LLM prompts.

competitive on raw accuracy, yet the per-class balance that Macro-F1 averages reveals a weak Healthy-detection arm. The other five models do not exhibit this asymmetry to the same degree, which suggests that LLaMA-3 under our prompt and shot configuration is unusually sensitive to the prior implicit in the 3-Healthy / 8-IBD demonstration mix. In a clinical screening context, where the costs of false-Healthy and false-IBD classifications are not symmetric [21], this is not a benign drift. Reporting accuracy alone would have placed LLaMA-3 above Mistral 7B in the LLM ranking, despite its substantially weaker discriminative balance—a methodological argument for retaining Macro-F1 as the primary metric whenever class prevalence departs from 50/50.

C. Feature Importance and Biological Consistency

The top-20 genera by Random Forest mean-decrease-in-impurity coincide exactly with the features supplied to all LLMs (Fig. 6). Roseburia ranked first (importance: 0.0356), Agathobacter second (0.0248), and Faecalibacterium third (0.0204)—all butyrate producers whose gut depletion in active IBD has been replicated across independent cohorts [13], [14]. The fact that the same three genera emerge consistently from 1,286 training samples suggests the supervised models picked up genuine biological signal rather than cohort-level artefacts, and that the biological prior encoded in the LLM prompts rests on the same statistically dominant features.

V. Discussion

We discuss the results along four axes: the proprietary vs. open-source performance gap, the role of prompt representation, the supervised vs. zero-annotation trade-off, and the limitations of the current benchmark.

A. Open-Source LLMs as Privacy-Preserving Alternatives

The best open-source model in this benchmark (Mistral 7B, Macro-F1: 0.5417) outperformed the best proprietary model (Gemini 2.5 Flash Lite, Macro-F1: 0.5333) by 0.0084 points, and the group-level gap of 0.039 Macro-F1 points falls below measurement uncertainty at $n = 30$. This is consistent with the practical use of locally-deployed open-source LLMs as zero-annotation classifiers in IBD microbiome contexts where commercial APIs are legally or logistically unavailable. The point is not that open-source models are uniformly better—they are not—but that the performance cost of local deployment appears small enough to be acceptable when labeled data are scarce. Any clinical application would require validation on a substantially larger independent cohort.

The within-open-source spread tells a different story: 0.130 Macro-F1 points separating Mistral 7B (0.5417) from DeepSeek-R1-Distill 1.5B (0.4118), which is more than three times the across-family gap. Model selection within the open-source ecosystem matters considerably more than the proprietary/open-source distinction itself. Mistral 7B and LLaMA-3 8B are both instruction-following models with straightforward chat templates; DeepSeek-R1-Distill is a reasoning model whose chain-of-thought generation consumed the token budget before producing the JSON output under standard inference conditions. This architectural difference—not necessarily any intrinsic capacity gap—largely explains the lower Macro-F1 of the DeepSeek variant tested here, and suggests that reasoning-oriented models may require different prompt designs (e.g., shorter context or explicit token budget constraints) to perform competitively on structured output tasks.

B. Representation as the Binding Constraint

The negligible inter-family gap is consistent with the broader finding of prior work: at 93% zero-inflation, prompt encoding drives more performance variation than model-specific capacity. In our preliminary encoding study, switching from P1 (raw counts, peak F1 = 0.46) to P3 (log-CPM top-10, peak F1 = 0.53) produced a larger performance shift than the entire proprietary vs. open-source gap. The finding that Mistral 7B—a 7B-parameter open-source model deployed at 4-bit precision—matches proprietary models with substantially larger parameter counts under P3 encoding further narrows the gap attributable to model size. This convergence supports and extends the findings of TabLLM [10] and LIFT [11] to the specific and more

TABLE V. HOLD-OUT PERFORMANCE (DATASET_30, $n = 30$). BOLD: BEST PER GROUP. CI: BOOTSTRAP 95% (1,000 RESAMPLES). DIFFERENCES BELOW 0.05 MACRO-F1 FALL WITHIN MEASUREMENT UNCERTAINTY

Model	Type	Features	Accuracy	Macro-F1	95% CI (F1)	AUC
Ensemble (RF+XGB+LGBM)	ML Classifier	799	**0.8333**	**0.7948**	[0.60, 0.94]	**0.7725**
XGBoost	ML Classifier	799	0.8000	0.7619	[0.55, 0.91]	0.7619
LightGBM	ML Classifier	799	0.8000	0.7443	[0.54, 0.91]	0.7566
Extra Trees	ML Classifier	799	0.7000	0.6306	[0.42, 0.81]	0.7249
Random Forest	ML Classifier	799	0.7000	0.5581	[0.38, 0.75]	0.7619
Gemini 2.5 Flash Lite	Proprietary LLM	20	0.5333	**0.5333**	[0.27, 0.63]	N/A
GPT-4o	Proprietary LLM	20	0.5667	0.4994	[0.33, 0.67]	N/A
Claude opus-4-5	Proprietary LLM	20	0.5333	0.4976	[0.32, 0.67]	N/A
				Proprietary group mean		*0.5101*
Mistral 7B	Open-source LLM	20	0.5667	**0.5417**	[0.37, 0.72]	N/A
DeepSeek-R1-Distill 1.5B	Open-source LLM	20	0.4643	0.4118	[0.32, 0.69]	N/A
LLaMA-3 8B	Open-source LLM	20	0.6333	0.4599	[0.32, 0.64]	N/A
				Open-source group mean		*0.4711*

demanding regime of zero-inflated compositional data, while going beyond those studies in providing a direct proprietary vs. open-source comparison. Future work evaluating larger open-source models (LLaMA-3 70B, Mixtral 8×7B) without quantisation would clarify whether the residual gap is genuine or an artefact of memory-constrained deployment.

C. LLM-Specific Evaluation Dimensions

Standard metrics only tell part of the story for language models. Consistency — how much predictions shift across runs with different seeds — needs at least three runs per model to measure. Calibration needs raw probability outputs over a larger sample, which we do not have here. Robustness to prompt variation needs a sweep across shot counts, encoding formats, and system message wording. Reasoning stability, mainly relevant for DeepSeek-R1, asks whether chain-of-thought traces land on the same label across calls. None of these were feasible without rerunning all six models under new conditions. We flag them here as the natural next step for anyone extending this benchmark.

D. Supervised vs. Zero-Annotation Performance

The supervised Ensemble outperformed the best LLM by 0.2531 Macro-F1 points (0.7948 vs. 0.5417), a gap driven by three structural asymmetries: annotation volume (1,286 training samples vs. 11 in-context demonstrations), feature access (799 genera vs. 20), and output calibration (probability scores vs. binary labels). This gap is expected and does not diminish the utility of the LLM approach in annotation-scarce settings. A Macro-F1 of 0.54 from 11 labelled examples may constitute a preliminary screening signal, though it falls well short of the performance thresholds typically required for clinical decision support—validating on a substantially larger and independently collected cohort would be a prerequisite for any clinical application [21].

E. Error Analysis and Clinical Interpretability

Across all six LLMs, the dominant error pattern on Dataset_30 is the misclassification of Healthy samples as IBD rather than the reverse. This false-positive bias is partially intentional—the asymmetric shot composition and the default-to-IBD rule in the system prompt were designed to prioritise sensitivity in a screening context—but it means that LLM-based classification in this setting would tend to over-refer rather than miss cases, which has different clinical cost implications depending on the use case [21]. The LLaMA-3 8B result discussed in Section IV-B is the most extreme instance: high accuracy coexists with near-complete failure to identify Healthy samples. From an interpretability standpoint, the top discriminant features (Roseburia, Agathobacter, Faecalibacterium) are biologically well-characterised butyrate producers whose depletion in active IBD has been replicated across independent cohorts [13], [14], meaning the classification signal reflects known disease biology rather than a statistical artefact. The threshold-based decision rules encoded in the prompt are derived from prior literature rather than fitted to the training data, however, and their validity in a new cohort would need to be confirmed before any clinical use.

F. Limitations

The holdout set contains only 30 samples, which constitutes the primary limitation of this study: differences below 0.05 Macro-F1 points are statistically uninformative at this sample size, and the conclusions regarding model ranking should be treated as preliminary observations. Validation on an independently collected cohort of substantially larger size would be necessary before any of these findings could inform a deployment decision. No batch correction was applied; the cross-validation to holdout AUC drop partially reflects technical variance that ComBat or MMUPHin normalisation [22] would partially resolve. Open-source models were evaluated at 4-bit NF4 quantisation due to GPU constraints (NVIDIA T4,

15 GB VRAM), which may introduce a precision disadvantage relative to proprietary models running at full precision on cloud infrastructure. LLM results reflect a single inference run at temperature $= 0.1$; residual stochasticity at low but non-zero temperature means that multi-run variance estimates (at least three independent runs) would be needed to confirm the robustness of individual model rankings. Running each model at least three times with different random seeds would be necessary to estimate within-model variance and to confirm whether the narrow performance differences observed—particularly between Mistral 7B and the proprietary models—are stable across runs. Until such multi-run analysis is conducted, individual model rankings should be interpreted with caution. The benchmark covers a single prompt configuration; shot count, example selection strategy, and encoding format remain unexplored axes of variation known to shift few-shot performance by non-negligible margins [10].

VI. Conclusion

The performance gap between proprietary and open-source LLMs in few-shot IBD microbiome classification is smaller than the field has generally assumed. When labelled data are available, the supervised Ensemble remains the stronger option (Macro-F1: 0.7948; AUC: 0.7725). Among the six LLMs tested under identical conditions, Mistral 7B—an open-source model running on local hardware—achieved the highest Macro-F1 (0.5417), ahead of GPT-4o, Claude, and Gemini. A mean inter-family gap of 0.039 Macro-F1 points does not justify exclusive reliance on commercial APIs in privacy-constrained clinical settings. The within-open-source spread of 0.130 points signals that model selection within the open-source ecosystem is the more consequential practical decision. Regardless of model family, log-CPM normalisation and zero-inflated feature elimination before serialisation remain the critical design choices for few-shot classification of sparse biomedical tabular data. Future work should evaluate unquantised open-source models of larger scale, incorporate explicit batch correction, conduct multi-run variance analysis to confirm LLM ranking stability, and extend this benchmark to autism spectrum disorder and colorectal cancer microbiome datasets to assess cross-disease generalisability.

Declaration on Generative AI

GPT-4o (OpenAI), Claude claude-opus-4-5 (Anthropic), and Gemini 2.5 Flash Lite (Google DeepMind) were used exclusively as experimental subjects in this study, evaluated under standard commercial API terms. No generative AI tool was used to write, edit, or improve any part of the manuscript text. All authors take full responsibility for the content of this publication.

Data Availability

The gut microbiome dataset is publicly available from the European Nucleotide Archive under accession numbers PRJEB13679 and PRJEB33711. Code, prompts, and reproducibility notebooks will be released at https://github.com/NouhailaEnnajih/LLM_ML_microbiome_IBD upon acceptance.

Acknowledgment

The authors declare no conflicts of interest. Mistral 7B, LLaMA-3, and DeepSeek-R1-Distill were used under their respective open-source licences. No promotional relationship exists with Anthropic, OpenAI, Google, Meta, Mistral AI, or DeepSeek.

References

[1] B. J. Callahan et al., "DADA2: High-resolution sample inference from Illumina amplicon data," *Nat. Methods*, vol. 13, pp. 581–583, 2016.

[2] GBD 2017 Inflammatory Bowel Disease Collaborators, "The global, regional, and national burden of inflammatory bowel disease in 195 countries and territories, 1990–2017," *Lancet Gastroenterol. Hepatol.*, vol. 5, pp. 17–30, 2020.

[3] J. Lloyd-Price et al., "Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases," *Nature*, vol. 569, pp. 655–662, 2019.

[4] E. Pasolli, D. T. Truong, F. Malik, L. Waldron, and N. Segata, "Machine learning meta-analysis of large metagenomic datasets," *PLOS Comput. Biol.*, vol. 12, p. e1004977, 2016.

[5] T. B. Brown et al., "Language models are few-shot learners," in *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877–1901, 2020.

[6] C. Duvallet, S. M. Gibbons, T. Gurry, R. A. Irizarry, and E. J. Alm, "Meta-analysis of gut microbiome studies identifies disease-specific and shared responses," *Nat. Commun.*, vol. 8, p. 1784, 2017.

[7] H. Nori, N. King, S. M. McKinney, D. Carignan, and E. Horvitz, "Capabilities of GPT-4 on medical challenge problems," *arXiv:2303.13375*, 2023.

[8] D. Van Veen et al., "Adapted large language models can outperform medical experts in clinical text summarization," *Nat. Med.*, vol. 30, pp. 1134–1142, 2024.

[9] K. Singhal et al., "Large language models encode clinical knowledge," *Nature*, vol. 620, pp. 172–180, 2023.

[10] S. Hegselmann et al., "TabLLM: Few-shot classification of tabular data with large language models," in *Proc. AISTATS*, vol. 206, pp. 5549–5581, 2023.

[11] T. Dinh et al., "LIFT: Language-interfaced fine-tuning for non-language machine learning tasks," in *Advances in Neural Information Processing Systems*, vol. 35, pp. 11763–11784, 2022.

[12] E. Triantafillou et al., "Meta-dataset: A dataset of datasets for learning to learn from few examples," in *International Conference on Learning Representations*, 2020.

[13] H. Sokol et al., "Faecalibacterium prausnitzii is an anti-inflammatory commensal bacterium identified by gut microbiota analysis of Crohn disease patients," *Proc. Natl. Acad. Sci. USA*, vol. 105, pp. 16731–16736, 2008.

[14] E. A. Franzosa et al., "Gut microbiome structure and metabolic activity in inflammatory bowel disease," *Nat. Microbiol.*, vol. 4, pp. 293–305, 2019.

[15] OpenAI, "GPT-4 technical report," *arXiv:2303.08774*, 2023.

[16] Anthropic, "Claude model family," https://www.anthropic.com/claude, 2026.

[17] Google DeepMind, "Gemini 2.5 Flash Lite model card," https://ai.google.dev/gemini-api/docs/models, 2026.

[18] Mistral AI, "Mistral 7B," *arXiv:2310.06825*, 2023.

[19] Meta AI, "LLaMA 3 model card," https://ai.meta.com/llama/, 2024.

[20] DeepSeek AI, "DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning," *arXiv:2501.12948*, 2025.

[21] D. T. Rubin et al., "ACG clinical guideline: Ulcerative colitis in adults," *Am. J. Gastroenterol.*, vol. 114, pp. 384–413, 2019.

[22] J. T. Leek, W. E. Johnson, H. S. Parker, A. E. Jaffe, and J. D. Storey, "The sva package for removing batch effects and other unwanted variation in high-throughput experiments," *Bioinformatics*, vol. 28, pp. 882–883, 2012.

[23] C. Liu, H. Han, Y. Qi, and W. Ling, "A knowledge-guided large language model framework for microbiome-based disease diagnosis," in *Proc. IEEE Int. Conf. Bioinformatics and Biomedicine (BIBM)*, 2025, pp. 7012–7018.

[24] J. Mu, Z.-Z. Tang, and G. Chen, "Systematic benchmarking of foundation models and classical baselines for microbiome-based disease prediction," *Research Square*, preprint, doi:10.21203/rs.3.rs-8912605/v1, 2026.

[25] J. Xing, H. Wang, Y. Sun, X. Su, and S. Wu, "Harnessing large language models to advance microbiome research: From sequence analysis to clinical applications," *Advanced Intelligent Discovery*, p. e202500038, 2025.

[26] B. Yan et al., "Recent advances in deep learning and language models for studying the microbiome," *Front. Genet.*, vol. 15, 2025, doi:10.3389/fgene.2024.1494474.

[27] C. K. Park et al., "Development of large language model specialized into microbiome datasets," *J. Microbiol. Biotechnol.*, vol. 36, p. e2511050, 2026.

[28] J. Cao, X. Xu, and X. Zhang, "Application of machine learning and large language model module for analyzing gut microbiota data," in *Advanced Intelligent Computing in Bioinformatics*, Singapore: Springer, 2024, pp. 37–48.

[29] H. El Massari, N. Gherabi, F. Qanouni, and S. Mhammedi, "Diabetes prediction using machine learning with feature engineering and hyper-parameter tuning," *(IJACSA) Int. J. Adv. Comput. Sci. Appl.*, vol. 15, no. 8, 2024, doi:10.14569/IJACSA.2024.0150818.

[30] S. Tomar, D. Dembla, and Y. Chaba, "Analysis and enhancement of prediction of cardiovascular disease diagnosis using machine learning models SVM, SGD, and XGBoost," *(IJACSA) Int. J. Adv. Comput. Sci. Appl.*, vol. 15, no. 4, 2024, doi:10.14569/IJACSA.2024.0150449.

[31] H. Almalki, A. O. Khadidos, and N. Alhebaishi, "Enhancing Alzheimer's detection: Leveraging ADNI data and large language models for high-accuracy diagnosis," *(IJACSA) Int. J. Adv. Comput. Sci. Appl.*, vol. 15, no. 11, 2024, doi:10.14569/IJACSA.2024.01511134.

[32] S. Li and X. Sun, "Dialogue-based disease diagnosis using hierarchical reinforcement learning with multi-expert feedback," *(IJACSA) Int. J. Adv. Comput. Sci. Appl.*, vol. 16, no. 2, 2025, doi:10.14569/IJACSA.2025.0160232.