A Binary Cascade Framework for Gut Microbiome Classification in Inflammatory Bowel Disease: Few-Shot LLMs, Co-Abundance Graphs, and Clinical Narrative Encoding

Nouhaila En Najih, Ahmed Moussa
Systems and Data Engineering Team-National School of Applied Sciences, Abdelmalek Essaadi University, Tangier, Morocco

***Abstract*—Subtyping inflammatory bowel disease (IBD) from gut microbiome sequencing data is a clinically demanding problem. Cross-cohort variability is large, and models trained on one dataset often fail on another. In this work, we propose a binary cascade pipeline that splits the three-class problem — healthy control, Crohn's disease (CD), and ulcerative colitis (UC) — into two independent binary classifiers optimized at each stage separately. A key component of the pipeline is a clinical narrative encoding strategy (P4) that converts centered log-ratio (CLR) microbial abundances into structured biomedical text, allowing a large language model (LLM) to apply its pre-existing biomedical knowledge without any fine-tuning. We evaluate Gemini-2.5-Flash under two encoding schemes, six few-shot sizes ($k \in \{2, 4, 6, 8, 10, 14\}$), and two example-selection strategies across three random seeds. The pipeline is benchmarked against a co-abundance graph neural network (GNN) and five supervised classifiers on an internal holdout set and two external cohorts: Dataset_30 ($n = 30$) and IBD423 ($n = 423$). P4 encoding achieves Macro-F1 of 0.539 ± 0.024 on Control vs. IBD and 0.585 ± 0.011 on CD vs. UC, without any gradient updates. On external CD vs. UC validation, Gemini reaches F1 = 0.873, matching the best supervised baseline while outperforming all tree-based and graph-based methods. Centroid-based few-shot selection outperforms random selection by 5.2 percentage points. The results indicate that grounding microbial measurements in clinical language is a practical route to interpretable, training-free IBD classification.**

***Keywords*—Inflammatory bowel disease; gut microbiome; large language models; few-shot learning; graph neural networks; 16S rRNA; clinical narrative encoding; binary cascade; centroid selection**

I. Introduction

More than 6.8 million people worldwide are affected by inflammatory bowel disease, and incidence is still rising in newly industrialized regions [1]. It has two subtypes: Crohn's disease (CD) and ulcerative colitis (UC). The two share several clinical features. However, they diverge considerably in where inflammation occurs, how tissue damage appears under the microscope, and how patients respond to treatment. Distinguishing the two subtypes, and ruling out healthy controls, is a precondition for any appropriate therapeutic plan. The gut microbiome contains a relevant disease signal. Studies conducted across independent cohorts have documented consistent reductions in butyrate-producing genera, particularly *Faecalibacterium prausnitzii* and *Roseburia*, alongside increases in pro-inflammatory taxa such as *Escherichia* and *Fusobacterium* [2]. Machine learning applied to 16S rRNA amplicon profiles has shown reasonable discriminative ability [3], [4]. However, two problems keep reappearing. First, classifiers trained on one cohort often break down on another: sequencing batch effects, differences in extraction protocols, and site-specific ecological drift all introduce variability unrelated to disease biology [5]. Second, high-performing ensemble methods are difficult to interpret, and without explainable reasoning, clinical adoption remains limited [6].

Large language models trained on biomedical text encode genus-level associations with disease states [7]. Whether that stored knowledge can substitute for labelled training data in a microbiome classification task is still an open question. A prior benchmark on the same dataset evaluated six frontier LLMs under fixed few-shot conditions with log-normalised features; Mistral 7B achieved the highest Macro-F1 at 0.5417, and open-source models remained competitive with proprietary APIs [8]. That study held the encoding format and example-selection strategy fixed. In this work, we vary both. Table I summarizes the differences between the two studies. We also include co-abundance GNNs, which operate on ecological interaction graphs rather than independent feature vectors and therefore capture structure that tabular models ignore [9], [10].

In this study, we evaluate LLMs, GNNs, and classical supervised learners on identical datasets under the same experimental protocol. Our contributions are as follows:

- A binary cascade decomposition that splits the three-class IBD problem into two sequential binary tasks, reducing per-stage class imbalance and allowing stage-specific model selection.
- A clinical narrative encoding (P4) that pairs each CLR-normalized genus abundance with a sentence drawn from the IBD literature, providing the LLM with both the measurement and its clinical meaning in the same prompt.
- A systematic few-shot ablation covering shot count ($k = 2$ to 14) and selection strategy (random vs. centroid-based), replicated across three independent seeds.
- A cross-paradigm comparison on two external cohorts that identifies where each model family generalizes and where it does not.

TABLE I. COMPARISON WITH PRIOR WORK [8]

Dimension	Prior work	This work
Focus	LLM benchmark	Encoding study
Prompt format	Fixed (P3 only)	P3 vs. P4
Few-shot count	Fixed k	$k \in \{2..14\}$ ablation
Example selection	Random	Random vs. centroid
Models compared	LLM only	LLM + GNN + ML
External valid.	1 cohort	2 cohorts
Multi-run stats	Single run	Mean $\pm$ std, 3 seeds

II. RELATED WORK

A. Supervised Learning on Microbiome Data

A meta-analysis of gut microbiome classification studies found that random forests achieve moderate accuracy within datasets but transfer poorly across cohorts [11]. The problem is that ensemble models fit sequencing platform signatures and extraction batch effects just as readily as genuine disease signal. Once applied to samples from a different centre, those cohort-specific artefacts vanish and predictive performance collapses. Logistic regression with ℓ_2 regularization has shown more stable cross-site behaviour; by shrinking coefficients toward zero, it avoids placing weight on narrow, site-dependent features [12]. More recent proposals include metaheuristic hyperparameter search for ensemble tuning [13] and recursive elimination of unstable microbiome features [4], [14]. Across all of these, the same trade-off appears: accuracy on the training cohort comes at the expense of accuracy on a new one. Our work asks whether this trade-off can be sidestepped by relying on prior biomedical knowledge rather than labelled cohort data.

B. Data Representation for Microbiome Classification

The choice of compositional transformation has a substantial effect on downstream results. Untransformed read counts confound library size with biological abundance. Rarefaction removes this confound but discards sequencing data. The CLR transformation handles both issues at once: it removes the compositional constraint while keeping all samples in the analysis, and it has been shown to reduce spurious genus-level correlations [15]. Feature selection is also important. The microbiome spans hundreds to thousands of genera, but only a fraction carry IBD-relevant signal. Random Forest importance scores, recursive elimination, and clinical prior knowledge are each valid selection criteria [16]. For LLM-based classification, which genera appear in the prompt is especially consequential since each genus occupies fixed context space. In this work, we combine data-driven importance ranking with mandatory inclusion of genera with documented IBD associations.

C. Deep Learning and Graph-Based Microbiome Analysis

Convolutional and transformer architectures applied to microbiome data have shown competitive in-distribution performance [3], [17]. Graph neural networks are well suited to this domain. Microbial communities are ecological networks, and co-occurrence and competitive relationships among taxa carry information that per-taxon abundance vectors do not capture [18]. Graph convolutional classifiers built on co-abundance networks have outperformed tabular baselines on several disease benchmarks [19], [20]. Whether this co-abundance structure itself transfers across sequencing sites, or whether it is as cohort-specific as the raw abundances, has not been systematically tested. In this study, we evaluate a co-abundance GNN on two independent external cohorts to provide direct evidence on this question.

D. LLMs in Biomedical Classification

In-context learning [21] has been evaluated across many biomedical tasks: medical question answering [7], [22], clinical note classification [23], and domain-adapted variants such as BioMistral [24]. The most directly relevant prior study benchmarked six LLMs on this same IBD dataset, finding open-source models competitive with proprietary APIs [8]. A separate study applied vision transformers to UC severity grading from endoscopic images [25]. None of these studies examined how the numerical encoding of omics inputs affects LLM performance. That is the gap this work addresses. Retrieval-augmented generation [27] takes a different approach: rather than fixing the in-context examples in advance, it draws them from a larger pool separately for each query. Centroid selection can be read as a simpler predecessor of this idea, since it selects representative examples once from the class means and reuses them across all queries. Per-query retrieval over a microbiome reference pool is examined in Section VI.

III. MATERIALS AND METHODS

A. Datasets

This study used three 16S rRNA sequencing datasets, with the breakdown of control and patient samples (CD and UC) presented in Table II.

TABLE II. DATASET SUMMARY

Dataset	Total	Control	CD	UC
Dataset_1 (primary)	1,286	336	731	219
Dataset_30 (Ext I)	30	10	10	10
IBD423 (Ext II)	423	44	204	175

Dataset_1 is the primary 16S rRNA amplicon cohort. We reserved a stratified holdout of 500 samples for Control vs. IBD evaluation and 300 samples for CD vs. UC, with the remaining samples forming the training pool. **Dataset_30** was collected at a separate sequencing centre and serves as a cross-site generalization benchmark. **IBD423** is a public dataset used only for external CD vs. UC validation.

B. Preprocessing

We rarefied all samples to 10,000 reads by multinomial subsampling and discarded samples below that depth. The CLR transformation was then applied:

$$\mathrm{CLR}(x_i) = \log\left(\frac{x_i + 0.5}{g(\mathbf{x} + 0.5)}\right), \tag{1}$$

where, $g(\cdot)$ denotes the geometric mean. A pseudocount of 0.5 was added to handle structural zeros.

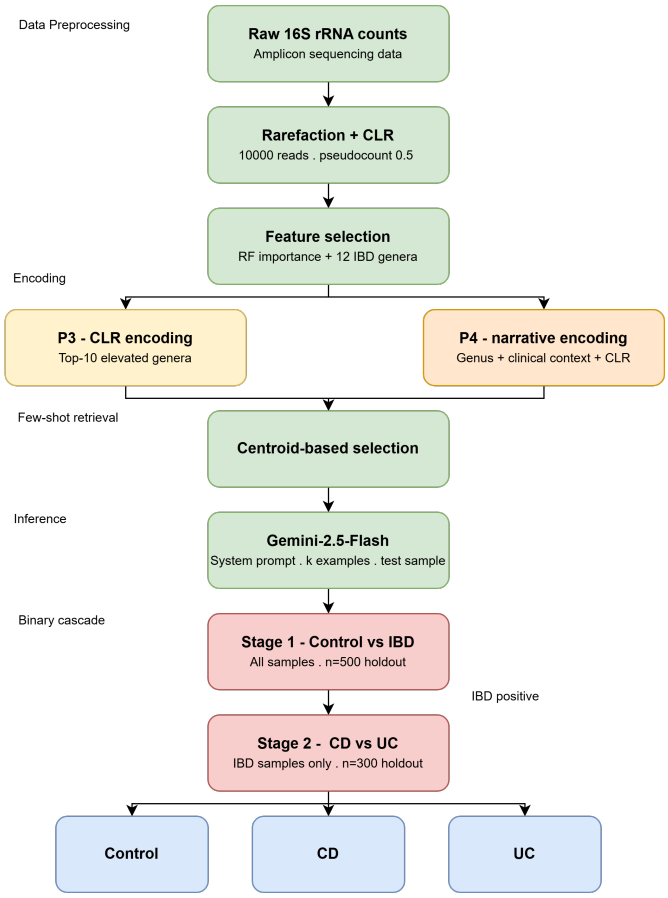

Fig. 1. Binary cascade pipeline. Stage 1 distinguishes healthy controls from all IBD patients. Stage 2, applied only to samples classified as IBD-positive, further distinguishes CD from UC. Each stage uses independent feature selection, encoding, and model training.

C. Feature Selection

For each binary problem, we fitted a Random Forest ($n_{\text{trees}} = 500$, max_depth $= 5$, balanced class weights) on the training pool and retained the top 20 genera by feature importance. We then appended twelve genera with documented IBD associations — including *Faecalibacterium*, *Roseburia*, *Akkermansia*, *Escherichia*, *Fusobacterium*, and *Ruminococcus* — if not already selected. The final feature sets contained 25 genera for Problem 1 and 26 for Problem 2.

D. Binary Cascade Framework

The three-class problem cannot be solved in one shot. We break it into two binary stages (Fig. 1). Stage 1 resolves a single decision: healthy or IBD. Samples it marks IBD-positive proceed to Stage 2, where the question becomes CD versus UC. The decomposition offers two advantages. Each stage faces a milder class imbalance than the full three-way split would impose, and because the stages operate independently, the encoding and the prompt can be tuned separately for the demands of each.

E. Microbiome Encoding Strategies

1) P3 - CLR abundance encoding: We select the genera with CLR values above the sample median (at most ten), apply ℓ_1-normalization, and format the result as a ranked list. No clinical context is added; the list communicates abundance magnitudes only.

2) P4 - Clinical narrative encoding: For P4, every selected genus is paired with a short literature-derived sentence stating its known IBD role, followed by whether it is elevated or depleted relative to the sample median and its exact CLR value. A representative entry reads:

```
Faecalibacterium prausnitzii is an
anti-inflammatory butyrate producer
depleted in IBD.
[Observed: depleted, CLR=-1.234]
```

The rationale is straightforward: the model receives the observation and its clinical meaning together, so it does not need to reconstruct the genus-disease link from pretraining memory under uncertainty.

F. Prompt Examples

The practical difference between the two encodings is easiest to see on a single sample. Table III places the P3 and P4 representations of the same CD patient side by side.

TABLE III. ENCODING COMPARISON FOR THE SAME SAMPLE (CD PATIENT). THE *Diagnosis:* FIELD IS LEFT BLANK TO MARK THE POSITION WHERE THE MODEL GENERATES ITS LABEL

P3 — CLR Encoding	**P4 — Narrative Encoding**
CLR normalized abundances: Faecalibacterium: 0.0312 Roseburia: 0.2841 Fusobacterium: 0.4102 Ruminococcus: 0.1745 ...	Clinical microbiome profile: *F. prausnitzii* is a butyrate producer depleted in IBD. [Observed: depleted, CLR=-1.23] *Fusobacterium* is enriched in CD. [Observed: elevated, CLR=+2.41] ...
Diagnosis:	*Diagnosis:*

P3 presents the same genera as bare CLR values, whereas P4 pairs each measurement with a sentence of clinical context before stating its observed direction and magnitude. The two columns encode identical numerical information; they differ only in how much biological meaning is made explicit to the model.

G. Few-Shot LLM Classification

1) Model and inference: We queried Gemini-2.5-Flash through the Google Generative Language API with temperature = 1.0 and maxOutputTokens = 100. A system prompt defined the clinical role and valid output labels. Each request included k in-context examples followed by the test sample. The model was instructed to output a single diagnostic label on the last line.

2) Example selection strategies: Both strategies selected $k/2$ examples per class.

a) Random selection: draws examples uniformly from the training pool under a fixed seed.

b) Centroid-based selection: computes the class centroid in CLR space and selects the $k/2$ training samples with the highest cosine similarity to that centroid, giving priority to the most representative profiles per class.

3) Shot count ablation: We tested $k \in \{2, 4, 6, 8, 10, 14\}$. All Experiment B runs used P3 encoding with centroid selection to isolate the effect of shot count from other variables.

H. Prompt Structure

To make the encoding concrete, it helps to see a full prompt exactly as the model receives it. Fig. 2 reproduces the complete input submitted to Gemini-2.5-Flash for the CD vs. UC task under P4 encoding, using $k = 4$ centroid-selected examples.

I. Co-Abundance Graph Neural Network

We computed pairwise Spearman correlations across all training-pool genera and connected genus pairs with absolute correlation above 0.3 by undirected edges. Node features were the per-sample CLR abundances. We trained a three-layer Graph Convolutional Network [20] with hidden dimension 64, dropout 0.3, and global mean pooling for 150 epochs using the Adam optimizer (lr = 10^{-3}, weight decay = 10^{-4}, batch size 32) with balanced cross-entropy loss. The checkpoint with the highest validation Macro-F1 was kept for evaluation.

J. Classical Supervised Baselines

We evaluated five classifiers: Extra Trees ($n = 500$, max_depth = 10), Random Forest ($n = 500$, max_depth = 10), Logistic Regression ($C = 0.1$, ℓ_2 penalty), XGBoost ($n = 200$, max_depth = 6, lr = 0.05), and LightGBM ($n = 200$, lr = 0.05). All models used balanced class weights and were trained on the full CLR feature matrix.

K. Evaluation Protocol

LLM experiments were run under three seeds (42, 123, 456). This controlled holdout subsampling (300 samples per run) and example pool composition. All LLM results are reported as mean $\pm$ std of Macro-F1 across the three seeds. GNN and supervised baselines were evaluated on the full holdout sets (500 samples for Problem 1, 300 for Problem 2) and on both external cohorts. LLM and supervised evaluations cover the holdout differently for reasons of cost. Each LLM prediction requires a paid API call, whereas supervised inference is effectively free once the model has been fitted. LLM subsampling was therefore kept class-stratified and repeated under three seeds, with results averaged across runs. Supervised baselines were evaluated on the full holdout sets, the setting in which they perform most favourably.

L. Computational Cost

Inference was performed with Gemini-2.5-Flash through the Google Generative Language API. A single request completed in one to two seconds and cost a small fraction of a US cent at the pricing tier used here. Across the full experimental grid of three seeds, six shot counts, two encodings, and two classification problems, total API charges came to approximately fifteen US dollars. GNN training completed in under 15 minutes on a T4 GPU, and each supervised baseline was fitted in under two minutes on CPU, both on freely available hardware. Extrapolating to deployment, a service processing tens to hundreds of samples per week would incur API costs in the low single-digit dollars per month.

[SYSTEM PROMPT]

```
You are a clinical microbiome specialist.
This sample is from an IBD patient.
Classify as:
- CD: Crohn's disease (depleted
  Faecalibacterium, enriched
  Fusobacterium/Ruminococcus)
- UC: ulcerative colitis (enriched
  Escherichia, mucosal inflammation
  pattern)
Output ONLY one word on the last line:
CD or UC
```

[FEW-SHOT EXAMPLES (k=4, centroid-based)]

```
Examples:
---
Profile:
Faecalibacterium prausnitzii is an
anti-inflammatory butyrate producer
depleted in IBD.
[Observed: depleted, CLR=-1.823]
Fusobacterium is enriched in CD.
[Observed: elevated, CLR=+2.341]
Diagnosis: CD
---
Profile:
Escherichia/Shigella is enriched in UC.
[Observed: elevated, CLR=+3.102]
Faecalibacterium prausnitzii is an
anti-inflammatory butyrate producer
depleted in IBD.
[Observed: depleted, CLR=-0.912]
Diagnosis: UC
---
```

[TEST SAMPLE]

```
Patient profile:
Faecalibacterium prausnitzii is an
anti-inflammatory butyrate producer
depleted in IBD.
[Observed: depleted, CLR=-2.104]
Ruminococcus is enriched in CD.
[Observed: elevated, CLR=+1.877]
Fusobacterium is enriched in CD.
[Observed: elevated, CLR=+1.543]
Diagnosis:
```

Fig. 2. Complete prompt for CD vs. UC classification under P4 narrative encoding with $k = 4$ centroid examples. The system prompt specifies the clinical role and output format. One example per class follows, each pairing a biomedical annotation with the observed CLR direction. The test sample appears last with the label blank.

IV. Results

A. Experiment A: Encoding Strategy

At a fixed budget of $k = 10$ centroid examples, P4 narrative encoding produces higher Macro-F1 than P3 CLR on both classification problems. The improvement is larger on CD vs. UC (+4.7 pp) than on Control vs. IBD (+0.7 pp). For the

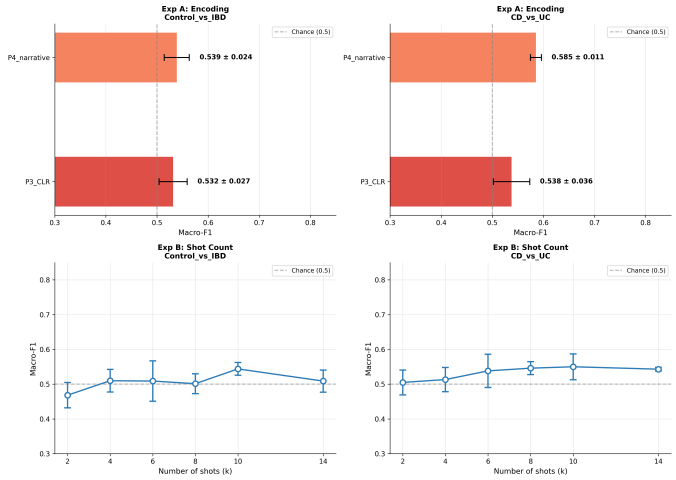

Fig. 3. LLM ablation results for Gemini-2.5-Flash. Upper panels (A, B): Macro-F1 comparison between P3 CLR and P4 narrative encodings with mean ± std over three seeds. Lower panels (C, D): Macro-F1 as a function of few-shot example count k using P3 CLR encoding and centroid-based selection.

subtype task, the clinical annotations attached to *Fusobacterium* (CD-associated) and *Escherichia* (UC-associated) give the model context it would otherwise need to recover from pretraining, and the results indicate that recovering it from memory is less reliable than stating it explicitly. Variance falls as well — from ± 0.036 under P3 to ± 0.011 under P4 on CD vs. UC — so P4 also produces more consistent results across different random example draws.

The two encodings are compared numerically in Table IV, averaged over 300 samples and three seeds, with the corresponding ablation shown in Fig. 3.

TABLE IV. EXPERIMENT A - ENCODING STRATEGY (MEAN ± STD, 3 SEEDS)

Problem	Encoding	Macro-F1
Control vs. IBD	P3 CLR	0.532 ± 0.028
	P4 Narrative	$\mathbf{0.539 \pm 0.024}$
CD vs. UC	P3 CLR	0.538 ± 0.036
	P4 Narrative	$\mathbf{0.585 \pm 0.011}$

B. Experiment B: Shot Count

We tested six shot counts ($k \in \{2, 4, 6, 8, 10, 14\}$) using P3 encoding with centroid selection. Results are summarized in Table V.

TABLE V. EXPERIMENT B — SHOT COUNT (MEAN ± STD, 3 SEEDS, P3 CLR)

k	Control vs. IBD	CD vs. UC
2	0.487 ± 0.018	0.497 ± 0.062
4	0.527 ± 0.039	0.536 ± 0.012
6	0.538 ± 0.048	0.579 ± 0.034
8	0.546 ± 0.019	0.546 ± 0.022
10	0.550 ± 0.037	0.555 ± 0.034
14	0.543 ± 0.006	0.559 ± 0.014

Performance improves from $k = 2$ to approximately $k = 6$–8, then levels off (Table V, Fig. 3). Beyond $k = 8$, additional examples do not appear to add useful distributional information. Including more examples past this point may actually dilute the most representative profiles with less typical ones. This pattern is consistent across both classification problems.

C. Experiment C: Example Selection Strategy

We compared centroid-based and random selection at $k = 10$, holding encoding (P3) and shot count fixed. Table VI reports the results across three seeds.

TABLE VI. EXPERIMENT C — SELECTION STRATEGY (MEAN ± STD, 3 SEEDS, $k = 10$)

Problem	Strategy	Macro-F1
Control vs. IBD	Random	0.487 ± 0.016
	Centroid	$\mathbf{0.539 \pm 0.035}$
CD vs. UC	Random	0.493 ± 0.012
	Centroid	$\mathbf{0.541 \pm 0.025}$

Centroid selection outperforms random selection by 5.2 pp on Control vs. IBD and 4.8 pp on CD vs. UC (Table VI, Fig. 5). Microbiome profiles are high-dimensional and sparse. A randomly drawn sample may sit far from the class centre of mass and give the model an atypical reference point. By selecting samples with high cosine similarity to the class centroid, centroid selection ensures the model receives archetypal rather than idiosyncratic examples.

D. Full Cross-Paradigm Comparison

We now compare all seven models across every evaluation set. Macro-F1 scores are reported in Table VII and visualized in Fig. 4.

On the internal holdout, tree-based ensembles perform best: Extra Trees and XGBoost both reach F1 = 0.73 on Control vs. IBD. Their external performance is very different. On Ext30, these same models fall to F1 = 0.33–0.44, a drop of nearly 30 percentage points. Ensemble methods trained on one sequencing site routinely pick up protocol-specific artefacts together with true disease signal; when those artefacts are absent in a new cohort, accuracy drops sharply [26]. Random Forest shows the same behaviour.

Logistic Regression departs from this pattern. Holdout accuracy is moderate (0.694 on Problem 1, 0.659 on Problem 2), yet both external cohorts remain competitive: F1 = 0.688 and 0.873 on Ext30, and F1 = 0.675 on IBD423.

On Ext30 Control vs. IBD, the GNN obtains its best result (F1 = 0.762), which suggests co-abundance graph structure captures disease-relevant ecological patterns that transfer across sites. For the CD vs. UC task, however, the same GNN drops to F1 = 0.333 on Ext30 — likely because of the smaller Stage 2 training set and the greater difficulty of subtype-level discrimination.

Without a single gradient update on the target data, Gemini-2.5-Flash with P4 encoding reaches F1 = 0.873 on Ext30 for CD vs. UC, tying Logistic Regression for first place and leaving all other models behind. On IBD423 it reaches F1 = 0.714, second

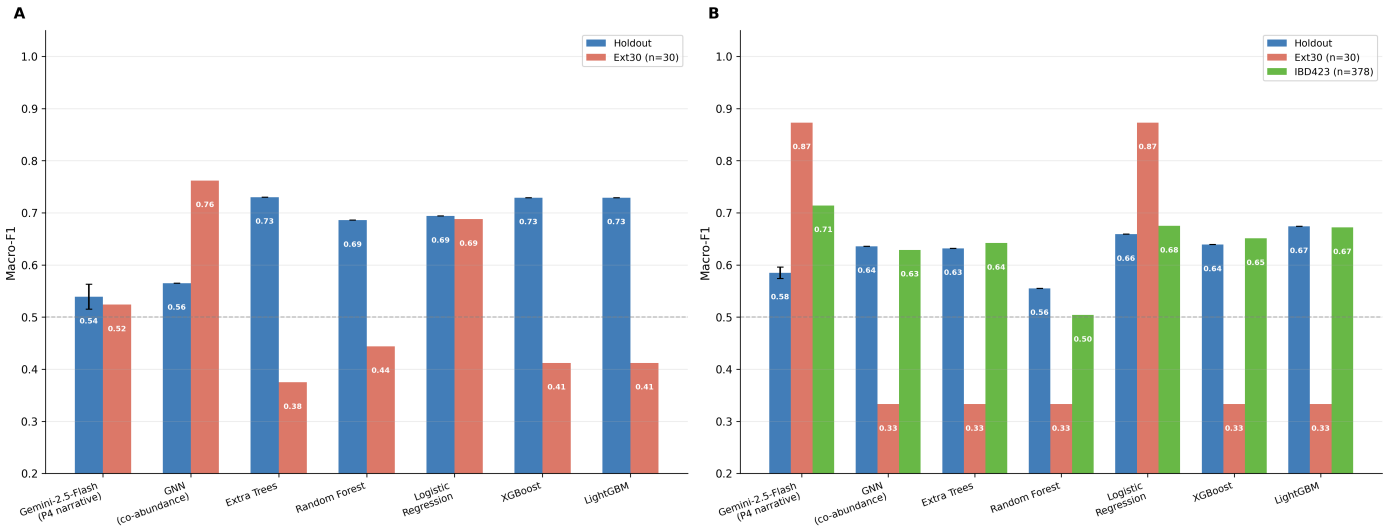

Fig. 4. Macro-F1 scores for all evaluated models on internal holdout and external validation cohorts. (A) Control vs. IBD classification. (B) CD vs. UC classification. Error bars on Gemini-2.5-Flash indicate standard deviation over three seeds. Dashed line marks chance level (F1 = 0.5).

TABLE VII. FULL MODEL COMPARISON - MACRO-F1 ACROSS INTERNAL AND EXTERNAL EVALUATION SETS

Model	Category	Control vs. IBD		CD vs. UC		
		Holdout	Ext30	Holdout	Ext30	IBD423
Gemini-2.5-Flash (P4)	LLM	0.539 ± 0.024	0.524	0.585 ± 0.011	**0.873**	0.714
GNN (co-abundance)	GNN	0.565	**0.762**	0.636	0.333	0.629
Extra Trees	ML	**0.730**	0.375	0.632	0.333	0.642
Random Forest	ML	0.686	0.444	0.555	0.333	0.504
Logistic Regression	ML	0.694	0.688	0.659	**0.873**	**0.675**
XGBoost	ML	0.729	0.412	0.639	0.333	0.651
LightGBM	ML	0.729	0.412	0.674	0.333	0.672

Bold: best result per column. LLM results are mean $\pm$ std over 3 seeds.

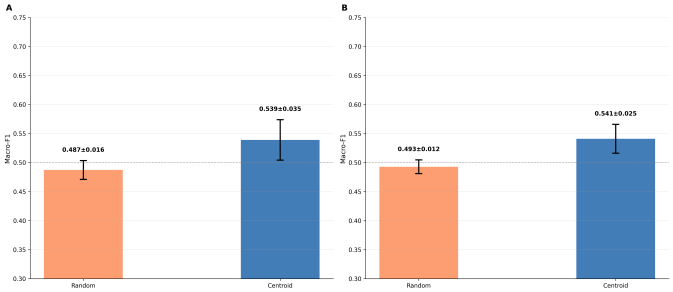

Fig. 5. Effect of few-shot example selection strategy on Gemini-2.5-Flash performance. (A) Control vs. IBD. (B) CD vs. UC. Error bars indicate standard deviation over three seeds. Centroid-based selection consistently outperforms random selection on both tasks.

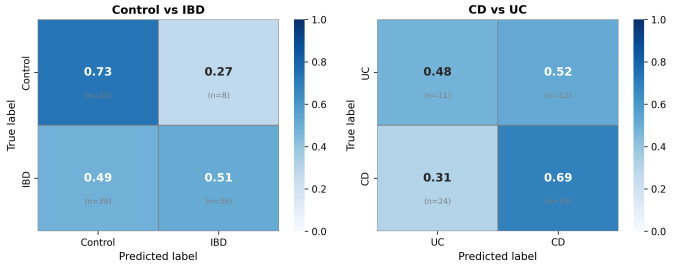

Fig. 6. Normalized confusion matrices for Gemini-2.5-Flash with P4 narrative encoding ($k = 10$, seed=42, $n = 100$ samples per problem). Values represent the proportion of true-class samples assigned to each predicted label; raw counts are shown in parentheses. UC samples are substantially harder to classify than CD samples (52.2% vs. 31.2% error rate), consistent with the greater microbiome heterogeneity of ulcerative colitis.

best on that cohort. P4 encoding and centroid selection are the only methodological differences between the two studies; consequently, the gain relative to [8] can be attributed to those two changes.

E. Confusion Matrix Analysis

Confusion matrices in Fig. 6 break down where each model succeeds and fails on each problem, using P4 encoding ($k = 10$, seed=42, $n = 100$ samples per problem).

Among healthy controls, 73% are correctly identified, while 48.6% of IBD samples are misclassified as healthy. The model seems to wait for a clear dysbiotic signal before assigning an IBD label, which pushes borderline samples near the boundary toward the healthy class. The split is wider for CD versus UC. Here the CD error rate is 31.2%, well under the 52.2% seen for UC. CD carries its own markers in the P4 prompt, chiefly the *Fusobacterium* and *Ruminococcus* annotations, while the UC annotations separate the classes far less cleanly. The recall

figures show the same gap.

V. Discussion

A. Clinical Narrative Encoding

The gap between P4 and P3 runs deeper than formatting. Consider a P3 entry like `Faecalibacterium: -1.823`. A genus and a number — nothing more reaches the model. For that pairing to carry meaning, the model must independently recall that this genus is anti-inflammatory and that its depletion marks active IBD — an implicit retrieval step with no guarantee of success. P4 removes that burden. Placed directly beside the measurement in the prompt, the clinical association ensures the observed direction and its biological meaning arrive together in a single token sequence. Rather than inferring what the number means, it applies context that is already present.

The two problems respond differently to the encoding change. On Control vs. IBD the gain is modest ($+0.7$ pp), while CD vs. UC shows a larger improvement ($+4.7$ pp). The reason is not surprising. Separating IBD from healthy controls rests on a broad, well-documented pattern: butyrate-producing taxa are depleted across the vast majority of IBD samples regardless of subtype, and that association is strong enough that a pretrained model recovers it without explicit annotation. Distinguishing CD from UC requires finer judgment — the relative elevations of *Fusobacterium* versus *Escherichia* are the key markers, and these are exactly the annotations that P4 supplies. In general, these findings suggest that LLM classification of structured omics data should be treated as a knowledge integration task. The prompt must bridge quantitative measurements and clinical significance; relying on the model to perform that bridge silently from pretraining is less dependable, especially for subtle distinctions. The mechanistic language used here describes patterns consistent with the observed outputs. A note on terminology is warranted here. Statements that the model "uses" or "applies" biomedical knowledge indicate only that the observed outputs are consistent with that interpretation. They are equally consistent with the model responding to surface features of the P4 text rather than to its biomedical content. Distinguishing between the two accounts would require ablation experiments in which portions of the prompt are removed or perturbed. Such experiments were not conducted here, and the question is revisited in Section V-F.

B. Error Analysis

In the Control vs. IBD task, misclassifications fall heavily on IBD samples labelled as healthy (34/70, 48.6%), whereas healthy controls are called correctly at a higher rate (8/30, 26.7%). Looking at the misclassified IBD samples, their *Ruminococcus* CLR is elevated relative to correctly classified cases (2.60 vs. 1.46), and their *Fusobacterium* CLR is lower (1.84 vs. 2.79). These profiles sit outside the canonical dysbiotic pattern that P4 annotations describe. This pattern is consistent with early-stage or inactive IBD, although it may equally reflect model sensitivity to CLR magnitudes near the annotation thresholds rather than any underlying biological distinction. Prediction outputs alone provide no basis for separating the two explanations.

For CD versus UC the error-rate gap is wide. UC samples were misclassified at 52.2% against 31.2% for CD. *Fusobacterium* CLR separated the two classes most cleanly, averaging 3.16 in correctly identified CD cases but only 1.27 in those that were missed. Once that genus stops being clearly elevated, the CD cue carried by the P4 prompt loses its grip and the sample drifts toward UC. Two other genera made things worse. Where *Blautia* (5.27 vs. 3.79) and *Faecalibacterium* (6.50 vs. 5.17) were both high, the butyrate-producing community still looked relatively intact, and that intact profile is where CD and UC proved hardest to separate. Widening the P4 annotation set with a few additional genera would be the next step to test.

C. Centroid Selection

A 5.2 pp gap between centroid and random selection is not trivial for a zero-training method. It reflects something specific about this data type. Microbiome profiles are sparse and high-dimensional, and a randomly chosen training sample can easily turn out to be an outlier, a patient whose profile shares the label but not the typical features. Hand that to the model as a reference and it learns the wrong prototype. Centroid selection sidesteps the problem by computing the class mean in CLR space and picking the samples closest to it. The examples that go into the prompt then look like the typical CD or UC profile. The computational cost is negligible. One matrix multiplication plus a sort over the training pool takes milliseconds, far less than a single API call. Under the conditions examined here, centroid selection emerges as the preferable strategy.

D. Model Selection Across Paradigms

No single model dominates every condition in Table VII; the results are better read as a deployment map than as a ranking. Tree ensembles suit single-site settings, where their high internal accuracy is an asset, but they degrade sharply once applied to external data. Logistic Regression is the conservative option for multi-site use, never leading in any condition yet never collapsing either. The GNN earns its place on Control versus IBD across sites (Ext30 F1 = 0.762) but not on CD versus UC, where its Stage 2 training set is too small and the task correspondingly harder. Gemini with P4 encoding is strongest in one specific configuration: cross-site CD versus UC discrimination with no local labels available.

These profiles translate directly into deployment guidance. A single-site pilot with annotated data is well served by Extra Trees or LightGBM. A multi-site rollout with mixed protocols should pair Logistic Regression at Stage 1 with the LLM at Stage 2, a combination the cascade architecture makes straightforward to implement.

Two considerations motivated the binary cascade: it reduces per-stage class imbalance, and it allows the encoding and prompt to be tuned separately for each decision. Alternative decompositions, including one-vs-rest schemes, hierarchical classifiers, and direct three-way prompting, were not benchmarked here. Comparing the cascade against these alternatives would clarify how much of the observed gain derives from the decomposition itself and how much from stage-specific optimization. The question remains open.

E. Practical Considerations

The LLM approach requires no labelled local training data at inference time. This is useful in clinical settings

where annotating a microbiome cohort is expensive and time-consuming. Since P4 already embeds clinical text in each prompt, the same inference step could be extended to generate natural-language justifications alongside the label, which would support clinician review without additional infrastructure. Data governance is a separate concern: sending patient microbiome data to an external commercial API raises privacy questions that the companion benchmark [8] addresses by evaluating locally deployable open-source models.

F. Limitations

LLM evaluations used subsampled holdout sets of 300 samples per run to control API costs, so the reported means may shift modestly at full-holdout scale. Ext30 contains 30 samples. At that size a single misclassification moves Macro-F1 by roughly three percentage points, and 95% bootstrap intervals would fall near ± 0.10. Results from this cohort are therefore best read as directional evidence rather than as a reliable ranking of models. IBD423, with $n = 423$, provides the more stable basis for assessing external performance.

The two cascade stages were executed independently, with ground-truth labels supplied to Stage 2. End-to-end error was consequently never measured. Multiplying the stage-level Macro-F1 values yields an approximate estimate of 0.32 on the internal holdout under an assumption of independent stage errors. Obtaining the true figure would require running Stage 2 on the IBD-positive predictions emitted by Stage 1, which is left for future work.

The encoding study varied two prompt dimensions: numerical representation (P3 versus P4) and example selection (random versus centroid). Instruction wording, example ordering, output constraints, chain-of-thought suppression, and context-length sensitivity were held fixed throughout. Given that prompt sensitivity in biomedical LLMs is well documented, a broader ablation across these dimensions would be a reasonable extension.

Results are reported as means and standard deviations over three seeds, without significance testing. Three seeds afford limited statistical power. Both headline differences exceed the observed spread by a wide margin: P4 over P3 by $+4.7$ pp on CD versus UC against a pooled standard deviation of 0.024, and centroid over random selection by $+5.2$ pp against a pooled standard deviation of 0.026. Paired testing across a larger number of seeds would nevertheless place these comparisons on firmer statistical ground.

Calibration, output entropy, and agreement across repeated LLM calls on identical prompts were not analysed. A temperature setting of 1.0 renders the model stochastic, so the same prompt can return different labels on consecutive calls. In a clinical context, per-sample variability and calibration quality matter as much as average Macro-F1, and that layer of analysis is absent from this work.

Several passages in the Discussion describe the model as engaging in biological reasoning. Such descriptions are interpretations of the outputs and should not be read as evidence. Establishing a causal claim would require targeted probing, for instance ablating individual annotations from the P4 prompt or perturbing single CLR values, which was not undertaken here.

P4 encodes curated clinical knowledge within each prompt, which means the experiment evaluates the model's capacity to apply written biomedical annotations to observed measurements rather than its ability to classify from numerical profiles without prior annotation. This scope should be kept in mind when interpreting LLM results alongside purely data-driven baselines.

VI. Conclusion

This study proposed a binary cascade framework for IBD microbiome classification and evaluated few-shot LLMs, co-abundance GNNs, and five supervised classifiers under identical conditions on an internal holdout and two independent external cohorts. P4 narrative encoding outperformed CLR-only encoding on both binary tasks, with the largest margin on CD vs. UC where genus-level annotations for *Fusobacterium* and *Escherichia* carried the most discriminative weight. Centroid-based selection outperformed random draws by approximately 5 percentage points on both problems, confirming that example representativeness is a consequential design choice for high-dimensional sparse microbiome data. Classification performance saturated at $k = 6$–8 shots. Tree-based ensembles achieved the highest internal accuracy but degraded substantially on external cohorts. Logistic Regression and Gemini-2.5-Flash with P4 encoding were the only approaches that remained competitive across all evaluation sets.

On cross-site CD vs. UC classification, the LLM reached F1 = 0.873 without task-specific training or local labels. In deployment scenarios where the target site differs from the training site and annotated data are unavailable, clinical narrative encoding offers a practical path to subtype-level discrimination when large labeled training cohorts are not available.

VII. Future Directions

Several directions follow from this work. Replacing the fixed centroid examples with dynamic retrieval over a large microbiome reference database, in the spirit of retrieval-augmented approaches [27], would allow each test sample to be matched against the most similar training profiles rather than the class mean, which should benefit atypical cases that sit far from the centroid. Training a language model on a targeted microbiome-disease corpus could eliminate the manual curation step by replacing hand-written P4 annotations with learned genus-level associations. Incorporating patient-level clinical variables — age, disease duration, medication history — into each cascade stage would bring the system closer to real deployment conditions; the two-stage architecture accommodates such extensions without structural modification. For supervised models, cohort-aware batch correction applied before the CLR step is a concrete intervention against the generalization gap observed here. A longer-term direction is combining the three model families: GNN structure for Control vs. IBD, LLM priors for CD vs. UC, and logistic regression as a stable baseline in multi-site settings, within a single ensemble that adapts its components to the deployment context.

Declaration on Generative AI

Gemini-2.5-Flash (Google DeepMind) was used exclusively as an experimental subject 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, PRJNA422193 and PRJEB33711. Code, prompts, and reproducibility notebooks are available at https://github.com/NouhailaEnnajih/IBD_LLM_microbiome_clinical_narrative_encoding.

Acknowledgment

The authors thank the contributors of the public IBD microbiome datasets used in this study.

References

[1] S. C. Ng *et al.*, "Worldwide incidence and prevalence of inflammatory bowel disease in the 21st century: a systematic review of population-based studies," *The Lancet*, vol. 390, no. 10114, pp. 2769–2778, 2017.

[2] F. Lloyd-Price *et al.*, "Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases," *Nature*, vol. 569, no. 7758, pp. 655–662, 2019.

[3] I. Manandhar *et al.*, "Gut microbiome-based supervised machine learning for clinical diagnosis of inflammatory bowel diseases," *American Journal of Physiology-Gastrointestinal and Liver Physiology*, vol. 320, no. 3, pp. G328–G337, 2021.

[4] Y. Lee, M. Cappellato, and B. Di Camillo, "Machine learning-based feature selection to search stable microbial biomarkers: application to inflammatory bowel disease," *GigaScience*, vol. 12, p. giad083, 2023.

[5] G. Papoutsoglou *et al.*, "Machine learning approaches in microbiome research: challenges and best practices," *Frontiers in Microbiology*, vol. 14, p. 1261889, 2023.

[6] Y. Peng *et al.*, "Comprehensive data optimization and risk prediction framework: machine learning methods for inflammatory bowel disease prediction based on the human gut microbiome data," *Frontiers in Microbiology*, vol. 15, p. 1483084, 2024.

[7] K. Singhal *et al.*, "Large language models encode clinical knowledge," *Nature*, vol. 620, no. 7972, pp. 172–180, 2023.

[8] N. En Najih, S. Hamida, and A. Moussa, "A controlled benchmark of open-source and proprietary LLMs for few-shot microbiome IBD classification," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 17, no. 5, 2026. DOI: 10.14569/IJACSA.2026.0170596.

[9] W. Ma *et al.*, "Deep learning in microbiome analysis: a comprehensive review of neural network models," *Frontiers in Microbiology*, vol. 15, p. 1516667, 2024.

[10] X. Song *et al.*, "Predicting microbe-disease associations via graph neural network and contrastive learning," *Frontiers in Microbiology*, vol. 15, p. 1483983, 2024.

[11] C. Duvallet *et al.*, "Meta-analysis of gut microbiome studies identifies disease-specific and shared responses," *Nature Communications*, vol. 8, no. 1, pp. 1–10, 2017.

[12] T. P. Quinn *et al.*, "Interpretable machine learning for genomics," *Human Genetics*, vol. 140, no. 9, pp. 1323–1363, 2021.

[13] A. Abd El-Aziz *et al.*, "Metaheuristic optimization techniques for machine learning in biomedical classification: a review," *Artificial Intelligence Review*, vol. 57, no. 4, 2024.

[14] E. Pasolli *et al.*, "Machine learning meta-analysis of large metagenomic datasets: tools and biological insights," *PLoS Computational Biology*, vol. 12, no. 7, e1004977, 2016.

[15] E. Ibrahimi *et al.*, "Overview of data preprocessing for machine learning applications in human microbiome research," *Frontiers in Microbiology*, vol. 14, p. 1250909, 2023.

[16] B. D. Topçuoğlu, A. Lesniak, M. A. Ruffin IV, J. A. Wiens, and C. Schloss, "A framework for effective application of machine learning to microbiome-based classification problems," *mBio*, vol. 11, no. 3, e00434–20, 2020.

[17] T. Nguyen *et al.*, "Deep learning methods for microbiome-based disease prediction: a systematic review," *Briefings in Bioinformatics*, vol. 25, no. 2, 2024.

[18] M. S. Matchado *et al.*, "Network analysis methods for studying microbial communities: a mini review," *Computational and Structural Biotechnology Journal*, vol. 19, pp. 2687–2698, 2021.

[19] Y. Huang *et al.*, "A graph neural network framework for microbiome-based disease prediction," *Briefings in Bioinformatics*, vol. 25,no. 2,2024.

[20] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2017.

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

[22] H. Nori *et al.*, "Capabilities of GPT-4 on medical challenge problems," *arXiv preprint arXiv:2303.13375*, 2023.

[23] C. Wu *et al.*, "LLMs-based few-shot disease predictions using EHR: a novel approach combining predictive agent reasoning and critical agent instruction," *AMIA Annual Symposium Proceedings*, pp. 319–328, 2024.

[24] Y. Labrak *et al.*, "BioMistral: a collection of open-source pretrained large language models for medical domains," *arXiv preprint arXiv:2402.10373*, 2024.

[25] D. Gupta, J. Gangrade, Y. P. Singh, and S. Gangrade, "A computer-aided diagnosis system for ulcerative colitis classification using vision transformer," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 16, no. 9, 2025.

[26] M. L. Calle, "Statistical analysis of metagenomics data: challenges and opportunities," *Frontiers in Microbiology*, vol. 14, 2023.

[27] P. Lewis *et al.*, "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Advances in Neural Information Processing Systems*, vol. 33, pp. 9459–9474, 2020.