ADAPTING LANGUAGE MODELS FOR MICROBIOME PREDICTION AND SCIENTIFIC REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

The human gut microbiome plays a vital role in shaping metabolism and immunity. To support research in this domain, AI systems must be able to reason about general microbiome knowledge and provide analysis of individual microbiome measurements, for example, predicting disease status from a sample. Current models are suboptimal: microbiome-specific models are inexpensive and can classify individual samples, but are not designed for open-ended knowledge tasks; frontier language models have scientific knowledge and few-shot classification abilities, but are expensive to run. In our work, we address this discrepancy by curating data to help inexpensive pretrained language models support both of these capabilities. Starting with sample classification, we introduce a continued-pretraining (CPT) dataset that combines information from sample metadata and scientific literature. We show that while current language models can be adapted for sample classification, training on our dataset improves performance. To evaluate microbiome knowledge, we introduce an expert-authored benchmark of challenging questions that require critical assessment of scientific situations. Finally, to support both sample classification and scientific knowledge, we curate a multi-task finetuning data mixture using data derived from biological databases, publications, and sample-level classification. A 1.7B model trained on this mixture outperforms few-shot frontier models by 7 percentage points on average on our sample classification tasks and outperforms a 31B model by 10 points on one of our two knowledge tasks. Our results illustrate the feasibility of curating data that supports predicting biologically meaningful sample properties and scientific knowledge for a complex scientific domain within a single, compact model.

1 INTRODUCTION

Foundation models have advanced biological sequence modeling (Hayes et al., 2025; Brixi et al., 2026), but supporting research on complex biological systems requires connecting quantitative measurements with existing scientific knowledge. The human microbiome, especially gut microbiota, provides a compelling setting: communities of gut microbes contribute to nutrient metabolism and modulate host immune responses (Sonnenburg et al., 2005; Caffrey et al., 2026, *i.a.*), and understanding these contributions at population scale requires linking quantitative community profiles to microbial functions and host physiology (Consortium, 2019; Lloyd-Price et al., 2019).

Supporting microbiome research therefore calls for complementary capabilities: predicting sample-level properties from community profiles, possessing knowledge about microbial and host biology, and critically assessing scientific claims and experimental designs. Existing microbiome models learn from community profiles and are commonly evaluated on classification tasks, including disease classification and identification of drug-treated samples (Pope et al., 2025; Zhang et al., 2026; Treloar et al., 2026); these models are generally not designed for open-ended scientific question answering. Conversely, general-purpose language models (LMs) can support scientific reasoning (Gemma et al., 2026; OpenAI, 2026), and can support classification through few-shot learning, but frontier models are expensive to run, and weaker models underperform compared to task-specific baselines. These complementary strengths and limitations motivate investigating whether we can curate data to make up the difference: to train inexpensive models that can support both profile classification and knowledge-intensive reasoning for microbiome research.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

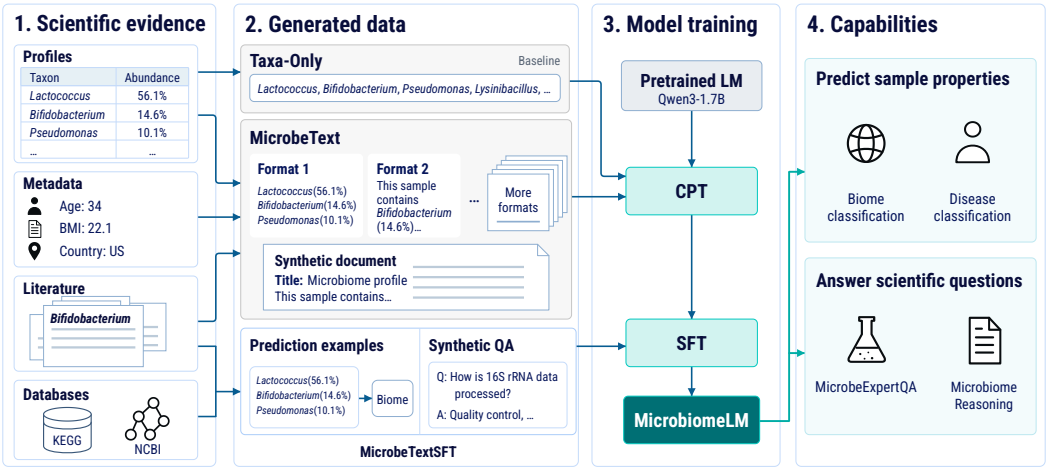


Figure 1: **Overview of our microbiome data-generation and model-training pipeline.** 1. Sample profiles, study metadata, literature, and biological databases provide scientific evidence. 2. `Taxa-Only` represents profiles as ranked taxon-name sequences. `MicrobeText` transforms study data and literature-derived facts into diverse profile formats and synthetic documents. `MicrobeTextSFT` mixes classification examples with synthetic question-answer pairs. 3. An existing language model is adapted through continued pretraining (CPT) and supervised finetuning (SFT). 4. Models are evaluated on sample-property prediction and scientific question answering.

In our work, we curate data to adapt pretrained language models to support prediction from microbiome profiles while having scientific microbiome knowledge (Figure 1). We begin with microbiome profile classification. Each sample is represented as an *abundance profile*, a text document listing the microbes present in a community and their relative abundances—in other words, which microbes there are and how common each is. Interpreting these profiles is central to microbiome research: they reveal community diversity, help track changes following treatment, and identify differences between healthy and diseased communities. This interpretation is challenging because profiles are complex: in our datasets, each profile contains approximately 100 microbes on average, drawn from approximately 15K distinct microbial types. Predicting properties such as disease status or drug response therefore requires learning how microbial identities and abundances jointly relate to these outcomes. However, these abundance tables can be formatted as text documents, making them accessible to LM-processing, rather than requiring domain-specific models.

To help LMs process abundance profiles, we introduce `MicrobeText`, a continued-pretraining (CPT) corpus that presents abundance profiles in diverse tabular formats and synthetic document styles, enriched with sample metadata and literature-derived information. Using few-shot prompting for frontier models and supervised finetuning for a smaller `Qwen3-1.7B` model, we establish that general LMs can perform several microbiome classification tasks. However, we find that using `MicrobeText` to continue-pretrain a `Qwen3-1.7B` model (`MicrobiomeLM-base`) improves mean macro-F1 across seven classification tasks from 66.8 to 70.0 under single-task finetuning and from 69.3 to 70.7 under multitask finetuning. Under both finetuning regimes, it achieves comparable or higher mean macro-F1 than a domain-specific model on five of the seven tasks. Ablations further show that the diverse tabular formats and synthetic documents are the main drivers of classification performance.

Turning to microbiome knowledge, we note the lack of existing benchmarks measuring scientific microbiome knowledge capability. As a result, we introduce `MicrobeExpertQA` (`MEQA`), an expert-authored benchmark of 50 long-form questions, each accompanied by a reference answer, supporting references, and a question-specific weighted rubric. The questions draw on practical research challenges, including compositional data interpretation, strain-level inference, experimental design, and the assessment of causal evidence. The rubrics reward scientifically justified answers and penalize predefined biological and methodological errors, allowing evaluation to account for both relevant knowledge and misleading interpretations.

Finally, to support both prediction and scientific knowledge within a single, inexpensive model, we construct a supervised finetuning (SFT) dataset, `MicrobeTextSFT`, which mixes classification examples with synthetic question-answer pairs created using biological databases and publications (Kanehisa & Goto, 2000). We use this dataset to train `MicrobiomeLM-MT`, a compact 1.7B, multi-task model that improves over the unadapted `Qwen3-1.7B` instruction-tuned baseline by 25.1 points on `MicrobeExpertQA` and 14.2 points on Microbiome Reasoning, outperforming `Gemma-4-31B-it` on the latter. These results demonstrate the feasibility of curating data to support quantitative prediction and scientific question answering within a single inexpensive language model, with minimal task-dependent trade-offs relative to task-specific models.

Our contributions are as follows:

- We curate `MicrobeText`, a CPT dataset that combines textual microbiome profiles, sample metadata, and literature-derived information to improve pretrained LM's microbiome classification ability.
- We introduce evaluation resources including `MicrobeExpertQA`, an expert-authored benchmark assessing microbiome knowledge, methodological reasoning, and critical evidence assessment.
- We curate `MicrobeTextSFT`, a dataset mixing classification tasks, microbiome knowledge, and synthetic scientific question-answer pairs. We use it to train a multitask model, `MicrobiomeLM-MT`, and quantify the benefits of domain CPT and the trade-offs of supporting both microbiome prediction and knowledge.[1]

2 Curating Microbiome Classification CPT Data

2.1 Dataset: MicrobeText

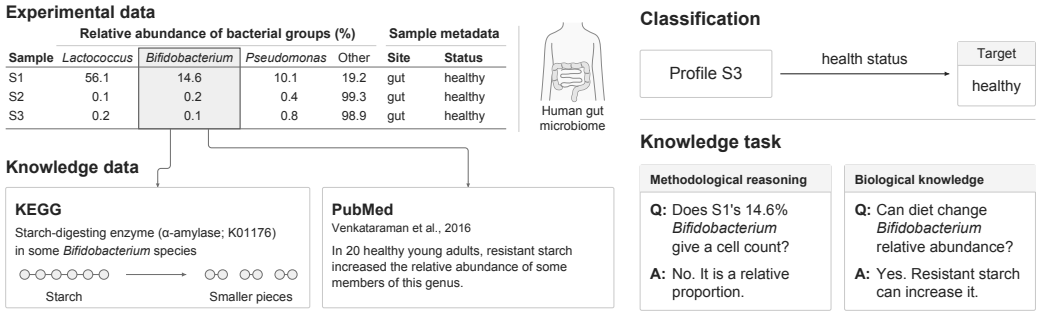
We now introduce `MicrobeText`, our continued-pretraining data, and study how it affects models' classification performance.

To curate continued-pretraining data that improves models' microbiome classification ability, we start from existing collections of abundance profiles. One collection is the Atlas dataset (Treloar et al., 2026), which contains 485K abundance profiles from the MGnify platform (Richardson et al., 2023), which aggregates microbiome data from sources like the European Nucleotide Archive. The second is the GMrepo database (Liu et al., 2026), which contains 60.2K additional samples aggregated from studies focusing microbiome-human disease interactions. We hold out 26.7K high-quality GMrepo samples for SFT and evaluation (Sec. 2.2), which leaves us with 33.6K high-quality samples for continued-pretraining. Each sample is an abundance profile with a list of the microbe's full taxonomic lineages and its relative abundance in the sample. Prior methods train from scratch on this data, so these collections are not designed with continued-pretraining in mind. To address this gap, we curate a dataset that (1) embeds abundance profiles into diverse contexts and (2) augments samples with linguistic information from scientific sources.

Diverse Contexts. Recent work on continued pretraining emphasizes that when teaching language models new information, it is important to incorporate that information into diverse contexts (Han et al., 2026; Lin et al., 2026). We incorporate diversity in two ways. First, for structural diversity, we generate **diversely formatted tables** by constructing formatters that insert the taxa into formats like Markdown tables, TSV files, and JSON objects. Second, we add diversity through the *document type* the table is inserted into. We generate fifty **diverse synthetic document** types (e.g. Wikipedia articles, Reddit thread forums, advertisements, legal document excerpts, etc.), and prompt a language model to rewrite the abundance profile into this document style (Prompts and examples in Appendix A.1).

External Knowledge. Sources like abundance profiles, metabolic pathways, sample metadata, and existing literature can intersect to improve language models' microbiome prediction abilities and knowledge. (See Figure 2(a) for an example.) As such, in addition to adding structural diversity,

[1] All data, models, and evaluations will be released upon publication.



(a) Data types (b) Task types

Figure 2: **Microbiome data and task types. (a)** Microbial names connect abundance profiles to structured biological annotations and scientific literature. **(b)** Classification tasks infer sample attributes from abundance profiles, while knowledge tasks assess biological understanding and methodological reasoning through scientific question answering.

we augment the abundance profiles with information from relevant scientific papers and sample metadata.

We obtain sample metadata from the MGNify API, including background information such as the title of the study that collected the sample and the location where the sample was taken (Richardson et al., 2023). Metadata is sample-specific but lacks information specific to the sample's microbes. To capture this, we query the PubMed search API with the microbial taxa to obtain relevant abstracts[2], and we prompt an LLM (`Gemma-4-31B-it`) to extract microbe-related information from the abstracts. We generate approximately 85K statements for 8K unique microbial genera from 90K PubMed abstracts. This gives us about one-fifth as many statements as samples, so to avoid overfitting to their surface forms, we randomly select 3-5 statements about bacteria found in the sample. For the diversely formatted tables, we append this additional information after the table to encourage the model to produce relevant information based on the table, similar to how the models will be evaluated. For the diverse synthetic documents, we incorporate this information into the synthetic document prompts.

Finally, the number of tokens is still relatively small, so we perform data augmentation on the abundance profiles themselves. We treat the profiles as probability distributions and sample from a Dirichlet distribution with parameters equal to the abundance values (Details in Appendix A.3). The end result is `MicrobeText`, which contains $\sim$2.3B tokens from diversifying the table structures (with each sample appearing five times) and $\sim$2.1B tokens from the synthetic documents (with each sample appearing once).

2.2 Evaluation

To evaluate the quality of `MicrobeText`, we CPT pretrained LMs and, treating them as microbiome sample representation models, evaluate them on classification tasks. Classification tasks require inferring a sample property from an abundance profile, such as predicting "healthy" disease status for sample S3 in Figure 2b. We adopt classification tasks from prior work in addition to proposing our own. From prior work, we use six tasks from the Compass benchmark that vary in their number of training samples and difficulties (Treloar et al., 2026). These tasks include identifying where in the human body or which stool-derived *in vitro* community (SIC) a microbiome sample was collected from, whether a sample was treated with a drug and if so what kind, and the age and birth mode of an infant a sample came from. (See Appendix C.1 for full descriptions).

The evaluation dataset that we introduce focuses on a task that is missing from these evaluations: disease-type classification. Our task leverages a 26.7K-sample subset of GMrepo data (Liu et al., 2026), with a train-test split of 25.9K/498/233. We clustered samples associated with diseases and healthy controls into eight categories for models to choose among, and ensured that any test samples

[2]We use abstracts because full texts are not always available.

from our pretraining data were omitted. Appendix C.2 provides more details and a breakdown of the disease type classes.

2.3 Experimental Setup

Prior work evaluates in a single-task training regime where a base model is finetuned separately on each task (Treloar et al., 2026), so we also adopt this approach.

Models. We compare four models. First, to test if language models can be adapted for microbiome classification, we compare to representative domain-specific model from prior work—`Waypoint-170M`, a 170 million-parameter model with a GPT-2 architecture trained on the Atlas dataset (Treloar et al., 2026). We choose this model because, compared to other prior works, it is larger in scale, has been trained on more samples, and performs well on the Compass evaluation task suite. Next, we compare to LM baselines. To see whether `MicrobeText` helps over no CPT at all, we finetune `Qwen3-1.7B` without CPT (**`No-CPT`**). To see whether the external knowledge and diverse contexts provide helpful supervision, we CPT on data without them (**`Taxa-Only`**). Specifically, we serialize abundance profiles using a strategy similar to (Treloar et al., 2026), converting each profile into a comma-separated list of microbial taxa ordered by their z-score-transformed abundance values. Finally, we compare these to **`MicrobiomeLM-base`**, a `Qwen3-1.7B` model trained on `MicrobeText`. When training `Taxa-Only` and `MicrobiomeLM-base`, 90% of the samples come from `MicrobeText` and the remaining 10% come from FineWeb-Edu to avoid catastrophic forgetting of language ability (Penedo et al., 2024). Hyperparameters are in Appendix E.

Single-task Finetuning. After the CPT step, we evaluate our models by performing single-task finetuning using the provided dataset splits. For the `Qwen` models, we perform finetuning by treating classification as a question-answering task: we insert the profile to classify and a short task description into the user message, and put the label into the assistant message (other hyperparameters in Appendix E). `Waypoint-170M` is trained differently: its vocabulary consists only of microbial taxa, so following Treloar et al. (2026), we pool the final hidden states and then train a classification head on them. This difference introduces an evaluation-time discrepancy: `Qwen` models have to choose the labels out of all possible next tokens, while the `Waypoint` models do not. To ensure a fair comparison, we evaluate the `Qwen` models using constrained decoding to force them to produce a valid label. We report the macro-average F1 for each dataset over three independently seeded runs.

Few-shot LMs Next, we investigate how well current LMs can classify microbiome profiles without parameter updates. To capture possible usage scenarios, we assess our base model, `Qwen3-1.7B` (with reasoning), a medium-sized open-weight model, `Gemma-4-31B-it` (with reasoning), and frontier proprietary LMs, `GPT-5.6-Terra` and `GPT-6-Astra` (reasoning = medium). We use five-shot classification to balance performance, coverage and cost considerations. (Additional analysis in Appendix D). Few-shot examples are chosen by finding the top 5 most similar training sample abundance profiles using Bray-Curtis dissimilarity (Bray & Curtis, 1957). In this setting, models also choose among valid labels using constrained decoding.

Classical Baselines. Prior work has found that classical machine learning approaches perform competitively with neural-based ones on microbiome tasks. As such, we implement and evaluate logistic regression, random forest, and XGBoost (Chen & Guestrin, 2016) models (hyperparameters in Appendix E).

2.4 Results

Table 1 shows the main results. We summarize the takeaways below.

Out-of-the-box LMs can be tuned to perform microbiome tasks. While the domain-specific `Waypoint-170M` model is successful on many tasks, `No-CPT` performs competitively in four of seven tasks (e.g., Drug/Not, Drug-class, Infant Age, and SIC). And compared to the logistic regression and random forest baselines, it performs the same or better in six of seven tasks. Regarding the few-shot LMs, `Gemma-4-31B-it` and `GPT-5.6-Terra` both achieve non-trivial performance

Table 1: Macro-F1 results on classification tasks. The models within a point of the best score are bolded. Datasets where our models are competitive or improve compared to `Waypoint-170M` are in italics. Numbers in parentheses are the standard deviations of three runs. Multi-task Training discussed in Sec. 3. **Takeaway: LM-based models perform comparably or better than domain-specific baselines for most classification tasks.**

	Biome Gut	SIC	Drug/Not	Drug Class	Infant Age	Birth Mode	Disease Clf
Classical ML Baselines							
Logistic Regression	79.6	89.7	63.2	16.1	71.8	68.4	48.9
Random Forest	62.2	84.7	78.3	17.0	69.5	68.8	44.4
XGBoost	80.8	94.9	**86.5**	**19.5**	75.8	72.1	65.0
LMs (5-shot)							
`Qwen3-1.7B`	7.0	79.2	32.1	0.9	9.3	30.7	13.6
`Gemma-4-31B-it`	67.5	70.6	77.4	12.2	59.4	63.4	17.3
`GPT-5.6-Terra`	57.1	86.0	75.8	*17.2*	54.4	64.5	51.3
`GPT-6-Astra`	73.4	89.7	77.5	*18.2*	63.5	71.8	52.2
Single-task Training							
`Waypoint-170M`	**95.1** (1.2)	93.4 (0.3)	80.0 (2.0)	15.9 (1.8)	71.9 (1.5)	**73.8** (1.5)	**66.5** (1.6)
`No-CPT`	70.2 (3.7)	***95.5*** *(0.3)*	*82.4* (1.3)	*17.2* (3.9)	*73.3* (3.8)	67.3 (3.5)	61.9 (4.0)
`Taxa-Only CPT`	86.2 (1.0)	*93.3* (2.7)	*82.0* (0.3)	*16.9* (1.2)	66.7 (4.4)	63.5 (5.5)	61.5 (3.2)
`MicrobiomeLM-base`	85.7 (1.6)	***95.1*** *(1.9)*	*85.0* (1.1)	*17.9* (1.7)	*72.1* (5.0)	67.9 (2.1)	***66.7*** *(2.1)*
Multi-task Training							
`No-CPT-MT`	85.0	*94.7*	*84.2*	***19.0***	*72.3*	71.3	58.1
`MicrobiomeLM-MT`	82.7	***96.0***	***85.6***	***19.7***	***77.0***	***73.8***	60.4

while `Qwen3-1.7B` fails. `GPT-6-Astra` outperforms logistic regression and random forest on four of seven tasks, and even `Waypoint-170M` in the Drug Class task.

Training with `MicrobeText` helps. Even though `No-CPT` can be finetuned to perform well on some benchmarks, performing CPT on microbiome data helps shore up the gaps. While both `Taxa-Only` and `MicrobeText` can help, `MicrobeText` helps more overall, particularly on Infant Age, Birth Mode, and our new disease classification task. Overall, `MicrobiomeLM-base` is competitive or better than the `No-CPT` model in six of the seven tasks, and competitive or better than `Waypoint-170M` in five.

Ablations. We investigate which features of our `MicrobeText` data drive the benefits we see over the `No-CPT` and `Taxa-Only` models. In three tasks, `MicrobiomeLM-base` clearly outperforms the `No-CPT` model, and in four `MicrobiomeLM-base` clearly outperforms the `Taxa-Only` model. We explore why by removing segments of the `MicrobeText` data and re-running CPT. First, we remove the synthetic documents to ablate the effect of including additional document structures (- synth docs). Then, we remove the external knowledge from the literature (- external). The results are in Table 2. On average, removing the synthetic documents causes a two-point drop, while removing the literature-based facts causes a further 0.8-point drop. The drop is most prominent in the Biome Gut and Disease Classification tasks. These tasks require predicting labels that are likely readily available in literature and sample metadata. However, when we remove literature facts, disease classification performance drops the most, suggesting that literature information helps associate diseases with microbiome samples.

Success of Simple Baselines. While our language model-based solutions are competitive with or better than `Waypoint-170M`, simpler models can yield competitive solutions on many classification tasks. In particular, we find that XGBoost outperforms many LM-based approaches. Others have noted XGBoost's relative strength in these tasks (Mu et al., 2026; Blekhman, 2026) and attribute this success to these methods' ability to handle sparsity and provide regularization for training on smaller datasets along with and ceilings on what can be predicted from profiles alone. While the XGBoost baselines are competitive for the single-task setting, they do not offer the flexibility of language models. If a task requires integrating language and microbiome samples, classical methods are not an option, and we can instead rely on methods like our multi-task models.

Table 2: Ablation results on individual representation learning tasks. Average scores on representation learning tasks illustrate that removing synthetic documents and the literature-derived facts from the curated MicrobeText data hurts performance.

CPT Data	Biome Gut	SIC	Drug/Not	Drug Class	Infant Age	Birth Mode	Disease Clf	Avg
No-CPT	70.2 (3.7)	**95.5** (0.3)	82.4 (1.3)	**17.2** (3.9)	**73.3** (3.8)	67.3 (3.5)	61.9 (4.0)	66.8
Taxa-Only	**86.2** (1.0)	93.3 (2.7)	82.0 (0.3)	**16.9** (1.2)	66.7 (4.4)	63.5 (5.5)	61.5 (3.2)	67.2
MicrobeText	**85.7** (1.6)	**95.1** (1.9)	**85.0** (1.1)	**17.9** (1.7)	72.1 (5.0)	67.9 (2.1)	**66.7** (2.1)	70.0
(- synth docs)	78.8 (1.0)	94.2 (1.5)	**84.3** (1.0)	**17.3** (3.3)	70.6 (0.8)	68.2 (2.1)	63.5 (2.5)	68.1
(- external)	79.4 (1.5)	94.3 (0.7)	**84.3** (1.8)	16.1 (2.8)	70.8 (4.0)	**69.9** (1.9)	56.1 (1.3)	67.3

3 Curating Multi-task SFT Data

Base models that can perform classification tasks alone are not sufficient to support microbiome research. Models must also have knowledge of the microbiome and the ability to reason about scientific processes (for examples of knowledge tasks, see Figure 2b). Unlike Waypoint-170M or other dedicated microbiome-based encoders, LMs can elegantly handle multitask scenarios due to their tokenization strategies and world knowledge. While the number of tasks we have available to train on in the microbiome domain is small compared to the number of tasks explored in frontier language SFT settings (Lambert et al., 2025), we use those that we have to train a proof-of-concept multi-task microbiome model by curating and finetuning on MicrobeTextSFT, a mix of microbiome classification and microbiome knowledge tasks.

3.1 Dataset: MicrobeTextSFT

We curate MicrobeTextSFT by mixing data to support microbiome classification and knowledge together. To support classification, we combine the training splits of the Compass tasks along with our Disease Type classification task. To support microbiome knowledge, we construct a synthetic corpus of question–answer pairs that require microbiome expertise to answer. We generate these pairs by combining knowledge from existing biological databases, including KEGG (Kanehisa & Goto, 2000), with information from PubMed publications and disease associations. We link database entries to relevant literature and disease context, then prompt a language model with this combined information to generate scientific questions and corresponding answers. Additionally, using a list of key concepts and microbiology ideas, we find abstracts that discuss these concepts and use them as seeds for generating $\sim$37K difficult microbiome questions. While small models have limited microbiome knowledge, larger models have more, so we distill answers to these questions from Gemma-4-31B-it. We also explore training on abstracts from PubMed publications but find that this does not provide much benefit, so we do not include them. Finally, we mix in 2K training samples used by (Lu et al., 2026), and 10K training samples from Tulu-v3 (Lambert et al., 2025) to encourage learning and preserving instruction-following formatting.

3.2 Multi-task Evaluation

To evaluate classification ability, we use the Compass and Disease Classification tasks; however, we perform no additional single-task finetuning. To assess microbiome knowledge, prior work is more limited. Traditional medical knowledge benchmarks contain few samples relevant to the human microbiome (See Appendix C.5). One exception is Lu et al. (2026), which curates the Microbiome Reasoning benchmark (MR) from the Microbiota-drug interaction and disease phenotype interrelation database (MDPID) (Yin et al., 2025). In addition to using this benchmark, we also curate our own.

We introduce MicrobeExpertQA, an expert-curated benchmark of 50 questions for evaluating microbiome knowledge and scientific reasoning through long-form question answering. Questions were written by expert microbiologists, drawing on practical challenges encountered in microbiome research and insights from relevant scientific and methodological literature. The questions emphasize research problems that require models to connect biological knowledge with data interpretation, experimental design, and critical assessment of evidence, instead of simply recalling isolated facts. Topics include compositional data analysis, strain-level inference, microbial ecology, causal

and functional validation, and microbiome-based interventions. For example, questions ask models to distinguish relative from absolute abundance, recognize when an assay lacks the resolution needed to support a claim, and identify critical experiments required to move from association toward causal evidence. Each question is accompanied by a detailed answer, supporting references, and a fine-grained, question-specific weighted rubric. Positive-weight criteria specify essential concepts, appropriate methodological choices, and relevant limitations, whereas negative-weight criteria penalize predefined scientific and methodological errors. This design targets not only the inclusion of relevant information but also the avoidance of misleading interpretations and unsupported conclusions. We use `Gemma-4-31B-it` as a judge to evaluate generated responses against these rubrics. (An example and judge prompt are in Appendix C.3.)

3.3 Multi-task Experimental Setup

We use `Qwen3-1.7B` and `MicrobiomeLM-base` as base models for multi-task training, training both on `MicrobeTextSFT` to obtain **`No-CPT-MT`** and **`MicrobiomeLM-MT`** respectively. Mixing the task data is not trivial—different tasks have different numbers of samples. Knowledge tasks also require much longer generations than classification tasks: the training data for knowledge tasks has 18.8M tokens, while the classification tasks ranged from 14K–406K tokens. To address this imbalance, we sample training samples with weights inversely proportional to the task's total number of tokens, and add extra weight to difficult classification tasks. For more details, see Appendix E. Finally, for comparisons to untuned LMs for knowledge tasks, no in-context examples are used.

3.4 Multi-task Results

The multi-task classification results are in Table 1 and knowledge results are in Table 3. On classification tasks, our multi-task models perform better than their single-task counterparts on some tasks (such as Drug class and Birth Mode), and worse on others (like Biome Gut and Disease classification). Improvements likely occur due to cross-dataset interactions—e.g., learning to distinguish which drug a sample was treated with likely helps distinguish those that were treated from those that were not—a phenomenon observed in prior multi-task learning settings (Ramsundar et al., 2015; Ruder, 2017). Finally, the gap between the `No-CPT-MT` and `MicrobiomeLM-MT` is visible in some datasets like infant age or birth mode, but largely disappears compared to the single-task setting, suggesting that training on a sufficient number of samples in the multi-task setting lessens the need for dedicated, domain-specific continued pretraining.

Table 3: Knowledge results. `MEQA` is `MicrobeExpertQA`, and `MR` is Microbiome reasoning. **Multi-task models trained on `MicrobeTextSFT` simultaneously can perform classification and knowledge tasks.**

	MEQA	MR
`Qwen3-1.7B`	7.12	57.0
`Gemma-4-31B-it`	51.2	61.8
`GPT-5.6-Terra`	**72.3**	73.4
`GPT-6-Astra`	71.8	**77.4**
`No-CPT-MT`	32.2	71.2
`MicrobiomeLM-MT`	33.2	71.6

In addition to helping with some microbiome classification tasks, the multi-task models also improve over untrained models in accessing microbiome knowledge. Compared to the untrained `Qwen3-1.7B` model, the multi-task models improve by 25–26 points on the scientific questions in `MicrobeExpertQA` and 14 points on the knowledge-based microbiome reasoning questions. We do not expect `MicrobeText` to provide much of a benefit in the knowledge elicitation setting, as most of the training data is dedicated to improving microbiome sample representations, but the fact that it performs approximately as well as the `No-CPT-MT` model indicates that it hasn't lost its linguistic ability during CPT. Compared to open-weight and frontier models, our 1.7B multi-task models underperform the `GPT` models on both tasks, but in the Microbiome Reasoning task, our model outperforms the `Gemma-4-31B-it` model.

4 Discussion

Value of Microbiome Data. Our results suggest that while current LMs can address the needs of microbiome research, training with additional domain-specific data is broadly helpful. For classification tasks, first training on `MicrobeText` usually provides a consistent improvement, though the boost shrinks during multi-task training due to the larger dataset size. This multi-task training

is also helpful—we don't observe a significant trade-off between single-task and multi-task performance on classification tasks. In fact, performance is higher across all but two tasks. These improvements provide evidence that frontier models' sub-optimal classification task performance can likely be overcome using datasets like the ones we present. Finally, on the knowledge tasks, `MicrobeTextSFT` can improve question-answering performance over the base models and larger open-weight models, but there is still room for improvement compared to the frontier models.

Clinical Applications. The tasks we study are only a starting point for multi-task microbiome modeling. For example, an important direction for future work is integrating longitudinal microbiome sampling with richer clinical data. Repeated microbiome measurements can be linked to longitudinal electronic medical record data, including medications, laboratory values, procedures, disease activity, and outcomes. Such datasets could allow models to move beyond static sample-level classification and instead learn clinically meaningful within-patient trajectories over time. Ultimately, this may enable models not only to predict disease states and trajectories, but also to identify microbiome-directed interventions that could alter those trajectories.

5 Related Work

Microbiome models. Our work fits into a broader set of microbiome profile works such as MGM, BiomeGPT, and Waypoint. Some train smaller models on more limited dataset, e.g., 1M parameter models on fewer than 20K samples (Pope et al., 2025; Medearis et al., 2026), while others train larger models, $\sim$8M parameters for MGM on $\sim$260K samples, and up to 170M parameters on $\sim$485K samples for Waypoint (Zhang et al., 2026; Treloar et al., 2026). Because these are microbiome-specific encoders, they tokenize abundance tables in different ways: usually each taxa gets its own token and abundance information is conveyed through order or additional abundance-specific tokens. In our work, we use pretrained LM's tokenizers without alteration. Despite these not being designed for microbiome tasks, we are still able to see comparable performance to prior work.

Integrating biological measurements with language. Much work in biological modeling, particularly for sequence data, trains domain-specific models, such as Evo 2 for DNA and ESM3 for proteins (Brixi et al., 2026; Hayes et al., 2025). Prior microbiome works we have discussed do similarly for microbiome community profiles. On the other extreme, some works focus predominantly on linguistic biological knowledge, training on biomedical literature for tasks like question answering (Luo et al., 2022; Bolton et al., 2024). In our work we argue for training models that can do both in the microbiome setting, as language models can treat both microbiome profiles and microbiome knowledge as text. In this way, our work aligns with Cell2Sentence wroks Levine et al. (2024); Rizvi et al. (2025), which train and finetune LMs to represent and reason over gene expression data; however, our setting is more data-limited with 100Ks of samples versus 10Ms, requiring additional augmentation. Another avenue toward integrating biological measurements with language involves using domain-specific sequence encoders as multi-modal inputs as done with audio or vision inputs for language models, rather than training unified models (Fallahpour et al., 2025; 2026). And a final approach is to delegate handling biological inputs and scientific knowledge to tools rather than processing them parametrically (Jin et al., 2024; Roohani et al., 2025). Closest to ours, Eubiota trains models to agentically reason and call tools that are relevant to microbiome queries (Lu et al., 2026); however these models are not trained on microbiome samples themselves.

6 Conclusion

Microbiome tasks are heterogeneous, but many inputs rely on language. Therefore, language models should form the underlying architectures for operating on microbiome data. In this work, we show that current language models can be adapted to perform well on microbiome classification tasks, and that additional continual pretraining on augmented microbiome data and training multi-task models can improve over domain-specific models. While they produce effective representations, these same models can also serve as sources of microbiome knowledge—answering factual questions and critiquing scientific workflows—and we introduce evaluations to study both aspects. The future of microbiome research lies in integrating various microbiome data sources into models, and the work we present here is a first step.

AI USE STATEMENT

In this work, we used generative AI tools to help generate synthetic datasets, for writing code, for suggesting experimental methods and parameters, for help formatting tables, and for finding relevant work during literature review.

For generating synthetic datasets, we used Claude Sonnet 4.8 to generate table formatting functions and to generate the list of synthetic document types. These were manually verified. The synthetic documents in `MicrobeText` were generated using `Qwen3-1.7B`, and the synthetic question-answer pairs in our microbiome knowledge SFT dataset were generated using `Gemma-4-31B-it`. These are large in number, so they were manually spot-checked rather than exhaustively verified.

For writing code, we used Claude Opus 4.8 and 5 and Claude Sonnet 4.8 to write pieces of code such as training configurations, new experimental conditions, and scripts for coordinating launching experiments. Claude products also wrote code for converting experiment results into LaTeX tables. All code was manually verified. For suggesting experimental methods and parameters, we used Claude Sonnet and Opus 4.8 to help debug issues with training hyperparameters.

Generative AI tools were not used in the writing of the submission (all em-dashes are our own), or for independently proposing, implementing, and running experiments. Nor were they used to interpret results or design the research methodology.

ETHICS STATEMENT

When working with microbiome data, it is important to protect the identities of any human subjects involved. None of the data we use comes from clinical sources that we collect—it is all from publicly available sources as collected by agencies like the European Nucleotide Archive. One ethical concern could be that, by pooling the data, we make it more accessible; however, the two sources of our data, Treloar et al. (2026) and Liu et al. (2026), already release microbiome data in aggregated form. Finally, and most importantly, all of the data we use is de-identified with no way to tie back to a particular human subject, and the subject metadata available is too coarse to identify an individual—for example country, age, sex, and disease status. Additionally, it is important to note that while our models provides predictions about properties like disease class, the models here are research prototypes and should not be used for patient diagnosis.

REPRODUCIBILITY STATEMENT

In order to ensure the reproducibility of our work, we take care to describe our methods in detail. The prompts we use for data curation and evaluation are included in the appendix, as are model training hyperparameters and more details about curation decisions. The code we use for data curation, model CPT and SFT, evaluation are also all available at the following link along with sample scripts for running them: `https://anonymous.4open.science/r/microbiome-iclr2027-submission-013D`. We plan on releasing all data, models, and code upon acceptance.

REFERENCES

Rahul K. Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, Johannes Heidecke, and Karan Singhal. Healthbench: Evaluating large language models towards improved human health. *ArXiv*, abs/2505.08775, 2025. URL `https://api.semanticscholar.org/CorpusID:278535396`.

Ran Blekhman. AI keeps failing at microbiome prediction. `https://blekhman.substack.com/p/ai-keeps-failing-at-microbiome-prediction`, March 2026. Ran's Substack (blog).

Elliot Bolton, Abhinav Venigalla, Michihiro Yasunaga, David Hall, Betty Xiong, Tony Lee, Roxana Daneshjou, Jonathan Frankle, Percy Liang, Michael Carbin, and Christopher D. Manning. BioMedLM: A 2.7b parameter language model trained on biomedical text, 2024. URL `https://arxiv.org/abs/2403.18421`.

J. Roger Bray and J. T. Curtis. An ordination of the upland forest communities of southern wisconsin. *Ecological Monographs*, 27(4):325–349, 1957. doi: https://doi.org/10.2307/1942268. URL https://esajournals.onlinelibrary.wiley.com/doi/abs/10.2307/1942268.

Garyk Brixi, Matthew G Durrant, Jerome Ku, Mohsen Naghipourfar, Michael Poli, Gwanggyu Sun, Greg Brockman, Daniel Chang, Alison Fanton, Gabriel A Gonzalez, et al. Genome modelling and design across all domains of life with evo 2. *Nature*, 652(8112):1349–1361, 2026.

Elisa B Caffrey, Elektra Kantzari Robinson, Jessica L Fessler, Anita Reddy, Samantha G Hernandez, Yifan Gao, Emma R Guiberson, Sean P Spencer, Erica D Sonnenburg, and Justin L Sonnenburg. Fermentation-derived metabolites shape host biology to attenuate severity of inflammatory and metabolic disease. *bioRxiv*, pp. 2026–07, 2026.

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016. URL https://api.semanticscholar.org/CorpusID:4650265.

The Integrative HMP Research Network Consortium. The integrative human microbiome project. *Nature*, 569(7758):641–648, 2019.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The faiss library. *IEEE Transactions on Big Data*, 12:346–361, 2024. URL https://api.semanticscholar.org/CorpusID:267028372.

Adibvafa Fallahpour, Andrew Magnuson, Purav Gupta, Shihao Ma, Jack Naimer, Arnav Shah, Haonan Duan, Omar Ibrahim, Hani Goodarzi, Chris J. Maddison, and Bo Wang. BioReason: Incentivizing multimodal biological reasoning within a DNA-LLM model. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-3728. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/a1fd4072fa16e13bcaed21823a03c695-Abstract-Conference.html.

Adibvafa Fallahpour, Arman Seyed-Ahmadi, Parsa Idehpour, Omar Ibrahim, Benedict MH Choi, Purav Gupta, Jack Naimer, Kevin Zhu, Abhinav Adduri, Arnav Shah, et al. Bioreason-pro: Advancing protein function prediction with multimodal biological reasoning. *bioRxiv*, pp. 2026–03, 2026.

National Center for Biotechnology Information (US). Entrez programming utilities help. Internet, 2010. URL https://www.ncbi.nlm.nih.gov/books/NBK25501/.

Gemma, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, Mayank Chaturvedi, Aditya Chawla, and et al. Gemma 4 technical report, 2026. URL https://arxiv.org/abs/2607.02770.

Seungju Han, Konwoo Kim, Chanwoo Park, Benjamin Newman, Suhas Kotha, Jaehun Jung, James Zou, and Yejin Choi. Synthetic mixed training: Scaling parametric knowledge acquisition beyond RAG, 2026. URL https://arxiv.org/abs/2603.23562.

Thomas Hayes, Roshan Rao, Halil Akin, Nicholas J Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q Tran, Jonathan Deaton, Marius Wiggert, et al. Simulating 500 million years of evolution with a language model. *Science*, 387(6736):850–858, 2025.

Rebecca Soskin Hicks, Mikhail Trofimov, D. Scott Lim, Rahul K. Arora, Foivos Tsimpourlas, Preston Bowman, Michael Sharman, Chio-Kin Tong, Kavin Karthik, Arnav Dugar, Akshay V. Jagadeesh, Khaled Saab, Johannes Heidecke, Ashley Alexander, Nate Gross, and Karan Singhal. Healthbench professional: Evaluating large language models on real clinician chats. *ArXiv*, abs/2604.27470, 2026. URL https://api.semanticscholar.org/CorpusID:287767203.

Qiao Jin, Yifan Yang, Qingyu Chen, and Zhiyong Lu. GeneGPT: augmenting large language models with domain tools for improved access to biomedical information. *Bioinformatics*, 40(2): btae075, 2024. doi: 10.1093/bioinformatics/btae075. URL https://doi.org/10.1093/bioinformatics/btae075.

Minoru Kanehisa and Susumu Goto. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30, 2000.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James Validad Miranda, Alisa Liu, Nouha Dziri, Xinxi Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Christopher Wilhelm, Luca Soldaini, Noah A. Smith, Yizhong Wang, Pradeep Dasigi, and Hannaneh Hajishirzi. Tulu 3: Pushing frontiers in open language model post-training. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=i1uGbfHHpH.

Daniel Levine, Syed Asad Rizvi, Sacha Lévy, Nazreen Pallikkavaliyaveetil, David Zhang, Xingyu Chen, Sina Ghadermarzi, Ruiming Wu, Zihe Zheng, Ivan Vrkic, et al. Cell2Sentence: teaching large language models the language of biology. *BioRxiv*, pp. 2023–09, 2024.

Jessy Lin, Vincent-Pierre Berges, Xilun Chen, Wen tau Yih, Gargi Ghosh, and Barlas Oguz. Learning facts at scale with active reading. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=mRi2cJDtIS.

Can Liu, Xiaohan Wang, Zhilin Zhang, Wenboxin Wang, Tianci Wang, Yuchen Zhao, Mingyu Wang, and Wei-Hua Chen. GMrepo v3: a curated human gut microbiome database with expanded disease coverage and enhanced cross-dataset biomarker analysis. *Nucleic Acids Research*, 54 (D1):D734–D742, 2026.

Jason Lloyd-Price, Cesar Arze, Ashwin N Ananthakrishnan, Melanie Schirmer, Julian Avila-Pacheco, Tiffany W Poon, Elizabeth Andrews, Nadim J Ajami, Kevin S Bonham, Colin J Brislawn, et al. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature*, 569(7758):655–662, 2019.

Pan Lu, Yifan Gao, William G Peng, Haoxiang Zhang, Kunlun Zhu, Elektra K Robinson, Qixin Xu, Masakazu Kotaka, Harrison G Zhang, Bingxuan Li, et al. Eubiota: Modular agentic AI for autonomous discovery in the gut microbiome. *bioRxiv*, pp. 2026–02, 2026.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6):bbac409, 2022. doi: 10.1093/bib/bbac409. URL https://doi.org/10.1093/bib/bbac409.

Nicholas A Medearis, Siyao Zhu, and Ali R Zomorrodi. Biomegpt: A foundation model for the human gut microbiome. *bioRxiv*, pp. 2026–01, 2026.

Ludovico Mitchener, Jon M. Laurent, Alex Andonian, Benjamin Tenmann, Siddharth Narayanan, Geemi Piyathma Wellawatte, Andrew D. White, Lorenzo Sani, and Samuel G. Rodriques. BixBench: a comprehensive benchmark for LLM-based agents in computational biology. *ArXiv*, abs/2503.00096, 2025. URL https://api.semanticscholar.org/CorpusID:276742500.

Jin Mu, Zheng-Zheng Tang, Guanhua Chen, et al. Systematic benchmarking of foundation models and classical baselines for microbiome-based disease prediction, February 2026. URL https://doi.org/10.21203/rs.3.rs-8912605/v1. PREPRINT (Version 1).

OpenAI. GPT-6 Astra System Card, September 2026. URL https://deploymentsafety.openai.com/gpt-6-astra/change-log.

Guilherme Penedo, Hynek Kydlíček, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, Thomas Wolf, et al. The fineweb datasets: Decanting the web for the finest text data at scale. *Advances in Neural Information Processing Systems*, 37:30811–30849, 2024.

Quintin Pope, Rohan Varma, Christine Tataru, Maude M David, and Xiaoli Fern. Learning a deep language model for microbiomes: the power of large scale unlabeled microbiome data. *PLOS Computational Biology*, 21(5):e1011353, 2025.

Bharath Ramsundar, Steven M. Kearnes, Patrick F. Riley, Dale R. Webster, David E. Konerding, and Vijay S. Pande. Massively multitask networks for drug discovery. *ArXiv*, abs/1502.02072, 2015. URL https://api.semanticscholar.org/CorpusID:2127453.

Lorna Richardson, Ben Allen, Germana Baldi, Martin Beracochea, Maxwell L Bileschi, Tony Burdett, Josephine Burgin, Juan Caballero-Pérez, Guy Cochrane, Lucy J Colwell, et al. Mgnify: the microbiome sequence data analysis resource in 2023. *Nucleic acids research*, 51(D1):D753–D759, 2023.

Syed Asad Rizvi, Daniel Levine, Aakash Patel, Shiyang Zhang, Eric Wang, Curtis Perry, Ivan Vrkic, Nicole Mayerli Constante, Zirui Fu, Sizhuang He, David Zhang, Cerise Tang, Zhuoyang Lyu, Rayyan Y. Darji, Chang Li, Emily Sun, David Jeong, Lawrence Zhao, Jennifer Kwan, David Braun, Brian P. Hafler, Hattie Chung, Rahul Madhav Dhodapkar, Paul F. Jaeger, Bryan Perozzi, Jeffrey Ishizuka, Shekoofeh Azizi, and David van Dijk. Scaling large language models for next-generation single-cell analysis. *bioRxiv*, 2025. URL https://api.semanticscholar.org/CorpusID:277940653.

Yusuf Roohani, Andrew Lee, Qian Huang, Jian Vora, Zachary Steinhart, Kexin Huang, Alexander Marson, Percy Liang, and Jure Leskovec. BioDiscoveryAgent: An AI agent for designing genetic perturbation experiments. In *International Conference on Learning Representations*, pp. 26417–26466, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/4252dc94531833029000f85dc5fac792-Abstract-Conference.html.

Sebastian Ruder. An overview of multi-task learning in deep neural networks. *ArXiv*, abs/1706.05098, 2017. URL https://api.semanticscholar.org/CorpusID:10175374.

Eric W Sayers, Jeffrey Beck, Evan E Bolton, J Rodney Brister, Jessica Chan, Ryan Connor, Michael Feldgarden, Anna M Fine, Kathryn Funk, Jinna Hoffman, et al. Database resources of the national center for biotechnology information in 2025. *Nucleic acids research*, 53(D1):D20–D29, 2025.

Justin L Sonnenburg, Jian Xu, Douglas D Leip, Chien-Huan Chen, Benjamin P Westover, Jeremy Weatherford, Jeremy D Buhler, and Jeffrey I Gordon. Glycan foraging in vivo by an intestine-adapted bacterial symbiont. *Science*, 307(5717):1955–1959, 2005.

Neythen J Treloar, Saif Ur-Rehman, and Jenny Yang. Learning the language of the microbiome with transformers. *bioRxiv*, pp. 2026–05, 2026.

Jiayi Yin, Hui Ma, Yuting Qi, Qingwei Zhao, Su Zeng, Fengcheng Li, and Feng Zhu. MDIPID: Microbiota-drug interaction and disease phenotype interrelation database. *Imeta*, 4(2):e70019, 2025.

Haohong Zhang, Yuli Zhang, Zixin Kang, Jiayun Xiong, Ronghua Yang, and Kang Ning. Mgm as a large-scale pretrained foundation model for microbiome analyses in diverse contexts. *Advanced Science*, 13(24):e13333, 2026.

Yu-Xin Zuo, Shang Qu, Yi-Fei Li, Zhangren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. MedXpertQA: Benchmarking expert-level medical reasoning and understanding. *ArXiv*, abs/2501.18362, 2025. URL https://api.semanticscholar.org/CorpusID:275993625.

A CURATING MICROBETEXT CPT DATASET

A.1 STRUCTURAL DIVERSITY.

We add structural diversity by inserting taxa into diversely formatted tables. We use Claude Code Sonnet to generate 42 table formatters that fall into three categories: lists of taxa, tables that include

Table 4: Sample categories and examples of diverse table formats.

Category	Count	Example
List	9	*Lactococcus* > *Bifidobacterium* > *Pseudomonas* . . .
Table	19	1. *Lactococus*: 56.1% 2. *Bifidobacterium*: 14.6% 3. *Pseudomonas*: 10.1% . . .
Buckets	14	Most Abundant: *Lactococcus*. Moderately abundant: *Bifidobacterium* and *Pseudomonas* . . .

taxa and abundance values, and bucketed lists, which group taxa by abundance without including the abundances. Examples are in Table 4.

Additionally, we add topical diversity by inserting the tables into one of 50 document types, generated by Claude Sonnet and manually verified. These cover formats such as Wikipedia-style entries, blog posts, social media threads, legal documents, among others. We prompt `Qwen3-1.7B` to incorporate the abundance table information into documents of those types. We choose `Qwen3-1.7B` for efficiency reasons, as we generated documents for all Atlas training samples. The prompt is below:

Synthetic Documents Prompt

```
Write a document that includes the information from the following microbiome abundance table:

{table}

Your document must include all bacteria and the bacteria should be in the same order as the ones in the table.

Here is some additional information that you can draw upon when creating your document:
Sample information:
{sample_info}

Relevant Publication(s):
{ publication_info }

Some additional bacteria facts:
{bacteria_facts}

Finally, your document must be formatted in a particular style:
{format_instructions}

REMEMBER:
- Your document must include all the bacteria and the bacteria should be in the same order as the ones in the table.
- Your document should be self contained. (e.g. don't mention "the table" without reproducing it or its contents. Don't use vague or
referential terms like "this study" without describing it first).
- Do not use markdown in your answer.
- Do not add any additional bacteria names.
- Do not provide any commentary commentary about the prompt or instructions.
```

And here is an example:

Example Synthetic Document (Style: Obituary)

In the intricate tapestry of the human microbiome, each bacterial species weaves its unique thread into the fabric of health and disease. This obituary honors the presence of these microorganisms, which, in their abundance, reflect the dynamic interplay between the host and the environment. The microbiome, a complex ecosystem that thrives on the skin's surface, is a testament to the resilience and adaptability of life , particularly in the context of the human body's most intimate and vital organ – the skin.

The top quartile of the microbiome, rich in species like Staphylococcus and Bacteroidetes, reflects a robust and diverse community that plays a crucial role in maintaining the skin's barrier function and immune response. These bacteria, often the first to colonize the skin, form a protective layer that shields the body from pathogens and environmental stressors. Their presence is a sign of a well–balanced microbiome, one that supports the host's health and resilience.

As we move down the list, the second quartile includes a mix of bacteria that contribute to the skin's metabolic processes and immune modulation. From the opportunistic pathogens like Burkholderiaceae to the beneficial microbes such as Lactobacillus, each species adds a unique dimension to the microbiome's functionality. These bacteria not only help in the synthesis of essential compounds but also in the regulation of the immune system, ensuring a harmonious relationship between the host and its microbial inhabitants.

The third quartile , with its diverse array of microorganisms, includes both beneficial and potential pathogenic species. Bacteria like Veillonella and Streptococcus, while often associated with health, can also contribute to the development of chronic conditions if their balance is disrupted. The bottom quartile, though less abundant, is equally important, with species such as Deinococcus and Solirubrobacter playing roles in the skin's defense against environmental stressors and pathogens.

In the context of the human body, the microbiome is a living entity , constantly adapting to the host's needs and the environment's challenges. The presence of these bacteria, in their respective abundances, underscores the importance of maintaining a balanced and healthy microbiome. Whether through proper oral hygiene, a diet rich in fiber, or the use of targeted therapies, the health of the microbiome is a reflection of the health of the host.

This obituary is not just a tribute to the bacteria, but a recognition of their vital role in the human body. Each species, in its unique way, contributes to the intricate web of life that sustains the skin and, by extension, the entire human organism. As we honor their presence, we also recognize the importance of preserving and nurturing this delicate balance, ensuring that the microbiome continues to thrive and support the health of the host.

A.2 External Knowledge.

The external knowledge we use when curating `MicrobeText` comes from two sources: sample metadata and scientific literature. When curating data from scientific literature, we query the PubMed search API with the bacteria names to obtain abstracts[3] from papers that are relevant to the bacteria in our samples (Sayers et al., 2025). The abstracts often focus on the overall scientific contribution rather than information about bacteria, so we extract factual information related to the bacteria by querying `Gemma-4-31B-it`. The data from GMrepo, which has disease labels, was collected first and is less numerous (33,569 samples). Because it was collected first, we search for `bacteria + disease + microbiome` and extract information one abstract at a time. For the Atlas-based samples, we batch the abstracts into groups of five before extracting the bacteria-specific statements. The prompt we use is below:

Scientific Literature Extraction Prompt

Summarize or extract a fact about the bacteria {bacteria} from the following scientific paper abstract.

Formatting instructions:
– your response should be a single sentence
– your response should be concise
– your response should not use markdown

{abstract}

A.3 Augmenting Abundance Tables.

Compared with other continued pretraining settings, the number of tokens is relatively small, so we augment the abundance tables themselves. We have no method to ensure the biological plausibility of novel abundance table compositions, so rather than generating them wholesale, we add noise to their relative abundance values to avoid models memorizing specific abundance values when they're present. Because the relative abundances have to be non-negative and add up to 1, we treat the problem of noising their values as sampling a probability distribution similar to the categorical distribution over the taxa in the samples. To do so, we sample from a Dirichlet distribution whose parameters are the existing relative abundances of the sample, $\vec{a}$: $\vec{a}' \sim \text{Dirichlet}(\gamma\,\vec{a} + \alpha)$. Doing so naively overly sharpens the distribution, so we use pseudocounts and a scale factor to flatten it.

[3] We use abstracts because full texts are not always available.

We heuristically determine that $\gamma = 500$ and $\alpha = 250$ provide a good distribution, and sample the noised relative abundances, $\vec{a}'$. Each atlas sample gives us five diverse table samples and one synthetic document sample because generating documents is expensive. This data augmentation results in a dataset of 4.5 billion tokens.

B CURATING MICROBIOME SCIENTIFIC REASONING SFT DATASET

We observe that an untuned `Qwen3-1.7B` model does not have much knowledge of microbiome scientific reasoning, so we design an SFT dataset to address this knowledge gap. Our SFT dataset is based on synthetic questions from PubMed Central publication abstracts. We use the NCBI Entrez Programming Utilities `pmc_bulk_fetch` tool to search the PubMed Central Open Access subset for papers related to the microbiome from 1990 until July 2026. Using the keyword `"microbiome"`, we are able to find 472,301 relevant full texts (for Biotechnology Information , US). We encode the paper abstracts using Qwen3-Embedding-0.6B and store them in a faiss index with L2 normalization Douze et al. (2024). Then, we generate a list of 69 microbiome-related concepts and query against this index for the top 300 abstracts for each concept. Next, we prompt `Gemma-4-31B-it` to generate microbiome-related questions conditioned on each abstract. Finally, while `Qwen3-1.7B` does not have much microbiome knowledge, `Gemma-4-31B-it` has more, so we distill the answers to these questions from `Gemma-4-31B-it`. At the end of this process, we end up with 37,235 question-answer pairs, using 36,735 for training and 500 for testing. The prompts for generating questions and answers are below:

SFT Question Generation Prompt

You are provided with the abstract of a scientific paper. Your task is to use the abstract as a jumping off point to generate expert–level question–answer pairs for the microbiome domain. Your questions should focused on practical methodologies and realistic decisions that microbiome scientists have to make when preparing their experiments and interpreting their results. Your questions should be complex. They should propose case studies and ask for critiques or interpretations. They should propose interpretations of experimental results, and ask questions about the correctness of those interpretations.
The basic structure of each question should be:

1. Introduce a scenario based on the abstract
2. Come up with a plausible conclusion that can be drawn from the abstract
3. Ask to evaluate either validity of the assumptions made, the quality of the analysis, etc. in an open–ended way.

Here's an example template that you can use for one question. (The others should be different.)
A collaborator performed [method] using [data] with dependent variable [y]. Because [complication], they [made the following corrections based on their knowledge and assumptions about the problem]. They want to report these as causal evidence. Evaluate the validity of this analysis and describe what would make the evidence credible.

Your should generate ˜3 questions in the following format:

Question: <question 1>
Answer: <answer 1>
–––
Question: <question 2>
Answer: <answer 2>
...

Notes:
– The question–answer pairs are going to be fed to a computer system independently. This means that they should not refer to each other in any way. They should not words like "abstract", "paper", "authors", etc. because the system will have no idea that they were based on a paper abstract
– The answers should be ˜a paragraph each. Aim for a few sentences.
– You should not use any additional markdown format other than separating the questions with "–––"
– the abstract should be used as inspiration for the questions–––you can ask about practical methodological decisions that the authors had to make, focus on alternative scenarios with different interpretations, or use the information in the abstract to propose a case study.
– Do NOT include factoid–style questions – the questions should represent realistic challenges that arise when conducting microbiome research. They should be difficult.

Here's the title and abstract to use for inspiration:

Title: {title}
Abstract: {abstract}

And here is the prompt for generating answers:

SFT Answer Generation Prompt

Answer the following question in detail. Keep in mind the following guidelines when generating the answer. You don't have to follow them exactly, but when appropriate you should cover everything they require:
1. **Check the prior.** Does the result contradict well-established biology (e.g., known dominant taxa in a well-studied population, known physiology)? A surprising result should raise suspicion of artifact before it is accepted as discovery.
2. **Interrogate the measurement before the biology.** Ask what the assay actually measures (DNA presence != viability; a spectral match != confirmed identity; gene annotation != enzymatic activity; relative abundance != absolute abundance != load).
3. **Look for confounding and directionality.** Could a third variable (diet, medication, collection site, batch, disease severity, transit time) explain both the "exposure" and the "outcome"? Could causation run backward?
4. **Distinguish correlation, association, and causal/mechanistic evidence.** What would it take to move from one to the next (isolation and culture, gnotobiotic/germ-free reconstitution, knockout/complementation, transfer experiments, randomized intervention)?
5. **Ask what would falsify the claim**, and whether that test was actually run.
6. **Propose the fix**, not just the critique: the right assay, the right control, the right statistical model, the right validation cohort.
Here's the question:

{question}

C Evaluation Details

Table 5 summarizes the tasks we study.

Table 5: The tasks we study. In addition to evaluating using tasks from prior work, Compass (Treloar et al., 2026), and (Lu et al., 2026), we introduce two new microbiome evaluation datasets.

Evaluation Type	Source	Dataset Name	Size (Train/Val/Test)
Classification	Treloar et al. (2026)	Gut Biome	6880 / 864 / 823
		SIC	2725 / 340 / 329
		Drug/Not Drug	3168 / 396 / 396
		Drug Class	1425 / 182 / 171
		Infant Age	1625 / 203 / 203
		Birth Mode	1625 / 203 / 203
	Ours	Disease Type Classification	25935 / 498 / 233
Knowledge	Lu et al. (2026)	Microbiome Reasoning	2000 / 548 / 548
	Ours	`MicrobeExpertQA`	- / - / 50

C.1 Compass Tasks:

Specific datasets from Compass that we use:

- Biome Gut: Requires classifying what part of the human gut a sample came from. This is a multi-prediction task, where models make either one or two predictions: one for a more general location (an 11-way classification task) and one for a more specific location (a 12-way classification task). All samples have a general location, and some also have a more specific one. Following Treloar et al. (2026), we provide the gold number of labels at evaluation time. For autoregressive classification, we provide all 132 possible combinations of the possible labels for constrained decoding.
- Stool-derived *In-vitro* Community (SIC): Requires classifying which stool derived *in-vitro* community a sample originates from (12-way classification task).
- Drug/Not Drug: Requires classifying whether a sample was treated with a drug or not (binary classification).
- Drug Class: Requires classifying the ATC drug class that a sample was treated with (14-way classification).
- Infant Age: Requires classifying the age of an infant when the sample was taken (6-way classification).
- Birth Mode: Requires classifying whether an infant was born via vaginal birth or C-section (binary classification).

The prompts we use for evaluation are below. For few-shot contexts, the few shot examples are inserted before the test profile.

Biome Gut

Given the following abundance table for a sample, identify the environment that the sample came from. Here are the options for the different environments. Not all labels will use all hierarchy levels, but you should be as detailed as you can be:

|-- Ceca
|-- Fecal
|-- Large intestine
| |-- Fecal
| |-- Sigmoid colon
| |-- Cecum
|-- Oral
| |-- Saliva
| |-- Throat
| |-- Supragingival plaque
| |-- tongue dorsum
| |-- Subgingival plaque
| |-- buccal mucosa
|-- Intestine
|-- Stomach
| |-- Rumen
|-- Midgut
|-- Digestive tube
| |-- Cecum
|-- Hindgut
| |-- Rectum
|-- Oral cavity
| |-- Buccal mucosa
|-- Foregut
|-- Rumen

Here is the abundance table:
{profile}

SIC

Given the following abundance table for a microbiome sample, identify the stool derived in-vitro community that the sample is most likely to come from. The options are given below:

|-- SIC0_heme
|-- SIC0
|-- SIC5
|-- SIC4
|-- SIC8
|-- SIC7
|-- SIC6
|-- SIC3
|-- SIC2
|-- SIC1
|-- SICMD
|-- SICcip

Here is the abundance table:
{profile}

Drug / Not

Given the following abundance table for a microbiome sample, identify if the sample is likely to come from environment that was treated with a drug (control=talse) or a control environment that was not treated with a drug (control=true).
The two possible labels are:
|-- False
|-- True

Here is the abundance table:
{profile}

Drug Class

Given the following abundance table for a microbiome sample, identify the ATC class of the drug that the sample was treated with.
The possible classes are:
|–– L
|–– J
|–– P
|–– G
|–– A
|–– M
|–– H
|–– D
|–– N
|–– V
|–– S
|–– C
|–– R
|–– B

Here is the abundance table:
{ profile }

Infant Age

Given the following abundance table for a microbiome sample that was taken from an inf, child, or mother. Classify which age bucket the child falls into. The possible labels are:

|–– 3Y
|–– 5Y
|–– M
|–– 4M
|–– 12M
|–– B

Where B = New Born (1 week), 4M = 4 months, 12M = 12 months, 3Y = 3 years, and 5Y = 5 years, M = sample from Child's mother.

Here is the abundance table:
{ profile }

Birth Mode

Given the following abundance table for a microbiome sample that was taken from an infant, child, or mother. Classify whether the child was born via C–section or Vaginal birth:

|–– C–section
|–– Vaginal

Here is the abundance table:
{ profile }

C.2 Disease Type Classification.

Our disease type classification task dataset uses samples from GMrepo (Liu et al., 2026) as the basis for our disease classification task. GMrepo is a curated database of abundance profiles and disease phenotypes, containing 118,965 samples from 1172 projects. We filtered these samples to remove those labeled as low-quality, bringing us to 68,723 samples, with 250 unique disease labels. The data is very long-tailed, with most diseases being associated with fewer than 100 samples, and approximately 45% of the samples are Healthy controls. If we predict the disease labels directly, it would be difficult to obtain enough signal to distinguish between models, so we sort the data into coarser labels—grouping common diseases (with $>$400 samples) into eight high-level disease categories and subsampling the healthy controls. We choose diseases with $>$400 samples, and group them into eight high-level disease categories (See Table 6 for the constituent labels and their counts). We find five test samples that overlap with `MicrobeText` and remove them from `MicrobeText`.

The generation prompt is below:

Table 6: Disease categories for our disease classification task.

Disease Category	Train	Val	Test	Diseases
Cancers / Neoplasms	4057	75	29	Breast Neoplasms, Colorectal Neoplasms, Prostatic Neoplasms
Gastrointestinal Disorders	5880	104	59	Colitis, Ulcerative, Crohn Disease, Diarrhea, Gastroenteritis, Inflammatory Bowel Diseases, Irritable Bowel Syndrome
Genetic Disorders	522	13	9	Cystic Fibrosis
Healthy	5864	122	10	Health
Infectious Diseases	2582	51	19	COVID-19, HIV Infections
Kidney Disorders	1185	21	19	Kidney Failure, Chronic, Renal Insufficiency, Chronic
Metabolic Disorders	3188	69	49	Diabetes Mellitus, Type 1, Diabetes Mellitus, Type 2, Diabetes, Gestational, Non-alcoholic Fatty Liver Disease, Osteoporosis, Postmenopausal
Neurological Disorders	2657	43	39	Alzheimer Disease, Autism Spectrum Disorder, Multiple Sclerosis, Parkinson Disease

Disease Classification

Given the following taxonomic profile of the abundance of certain bacteria in a human gut sample, predict what category of disease the human has or say "Healthy" if there is no disease.

Taxonomic Profile (genus):
{ profile }

Here are the possible disease types to choose from:
– Cancers / Neoplasms
– Gastrointestinal Disorders
– Genetic / Congenital Disorders
– Healthy
– Infectious Diseases
– Kidney Disorders
– Metabolic / Endocrine Disorders
– Neurological / Neurodegenerative Disorders

C.3 MicrobeExpertQA Example and Judge Prompt

Below is an example from `MicrobeExpertQA` and the prompt we use to judge the answers. At evaluation time, the models are provided only with the question.

`MicrobeExpertQA` Sample Question and Rubric

Differential abundance from compositional amplicon data

In a longitudinal antibiotic –perturbation study, our 16S data show that the relative abundance of Bacteroides fragilis rose from 5% to 25% after a 5–day course of oral vancomycin. A colleague concludes that the antibiotic promoted B. fragilis growth. Is that interpretation justified , and how would you determine what actually happened?

Rubric:
– Recommends rarefying to even depth as the sole/primary normalization without acknowledging its information loss or the ongoing methodological debate. (–3.0)
– Recommends or applies standard parametric/nonparametric tests (e.g., t–test, Wilcoxon) directly on relative abundances or proportions without addressing compositionality.(–8.0)
– Explicitly states that amplicon data are compositional (constrained to a constant sum) and that this constraint invalidates naive correlation /abundance inference (e.g., induces spurious negative associations). (10.0)
– Recommends a compositionally–aware differential abundance method (e.g., ANCOM–BC, ALDEx2) or an explicit centered–log–ratio (CLR) transformation prior to testing. (8.0)
– States that the false discovery rate must be controlled across the many taxa tested (e.g., Benjamini–Hochberg). (5.0)
– Includes prevalence/abundance filtering of rare features and/or explicit handling of the zero– inflation /sparsity problem as preprocessing. (4.0)
– Notes that DA results are method–dependent and that agreement across multiple methods (or a sensitivity analysis) increases confidence. (3.0)

The prompt we use to evaluate the generated answers is below:

MicrobeExpertQA Evaluation Prompt

```
You are evaluating whether a model response correctly addresses a user's query in a microbiome context. You are provied with the model responses, a rubric item, and a sample answer (that satisfies the rubric item if its score is positive or omits the error if the rubric item is negative). Your job is to determine the extent to which the rubric item is *satisfied* or *not satisfied*.

# Query
{query}

# Sample Response
{sample_response}

# Model Response
{model_response}

# Rubric Item
{rubric_item}


# Instructions
Evaluate whether the model's response covers the error described in the rubric item. We only care about recall - it is fine if the model response identifies additional points not mentioned in the rubric.
Note: you should be strict with your scoring. A response should achieve a perfect score only if it includes everything in the rubric item. For example, consider the following:
(Part of a) Response: "### Step 1: Isolation and Cultivation (The "Bottleneck" Step)
* **Strategy:** Use **culturomics**. This involves high-throughput anaerobic culturing using a wide array of media, varying incubation conditions (pH, temperature), and utilizing "co-culture" techniques (growing the target species alongside known helper strains that provide essential growth factors)."

Rubric Item: [Possible points: 6] Notes that the organism must first be cultured/isolated (or a representative isolate obtained) given it is currently uncultured, and that this is a nontrivial, rate-limiting step (culturomics).

Here, the score should be 5. This is because while the part of the response mentions that the organism must be isolated and cultured, and that it is non-trivial, it does not mention that this step is rate-limiting.

Your final response should be formatted as JSON object with two fields:
- "explanation": a string explaining your reasoning
- "score": the number of points the model's response should obtain for this rubric item.

Return just the JSON object in markdown format. Do not include any other text.
```

C.4 Microbiome Reasoning

The prompt we use for microbiome reasoning evaluation is below.

Microbiome Reasoning Evaluation Prompt

```
SYSTEM MESSAGE:
You are a helpful and precise assistant for answering microbiome-related questions. The last line of your answer should be a single letter corresponding to the correct answer choice on its own line.

USER MESSAGE:
{question}
```

C.5 Microbiome Questions in Medical Benchmarks.

While in this work, the datasets that we are interested in dealt with scientific microbiome knowledge, there are additional, more general medical benchmarks as well. We conduct some preliminary analysis and determine that most modern medical benchmarks do not contain microbiome-related samples of sufficient number, difficulty, or diversity. We analyze HealthBench, HealthBench Professional, MedXpertQA, and Bixbench (Arora et al., 2025; Hicks et al., 2026; Zuo et al., 2025; Mitchener et al., 2025) and prompt `Gemma-4-31B-it` to filter samples from these datasets that require knowledge of the human microbiome. In HealthBench Professional, three samples pass the filter, 41 in MedXpertQA, and 7 in BixBench. HealthBench has the most with 72, though most are about general problems with the digestive system rather than specific to the microbiome.

D Few-shot Experiment Analysis

We analyze how different numbers of in-context examples affect our few-shot microbiome classification LM results. We test four models on 0, 5, 10, and 20 shots, with the results in Figure 3. In

general, we find that five in-context examples lead to a reasonable balance of performance, coverage, and cost. For most tasks, the largest jump in performance comes from shifting from zero examples to five, and the improvements after that are relatively small. Only in the Drug/Not Drug task for `Qwen3-1.7B` does using five in-context examples hurt. In terms of coverage, if we increase the number of in-context examples to 10, we observe context-length issues: namely, `Qwen3-1.7B` evaluated 10-shot on Biome Gut has some examples that exceed the context length, and context length also becomes an issue for Disease classification at 20 in-context examples. Finally, in terms of cost, running gpt-5-astra on all evaluations costs ∼170 USD, so five in-context examples provide a reasonable performance/cost ratio.

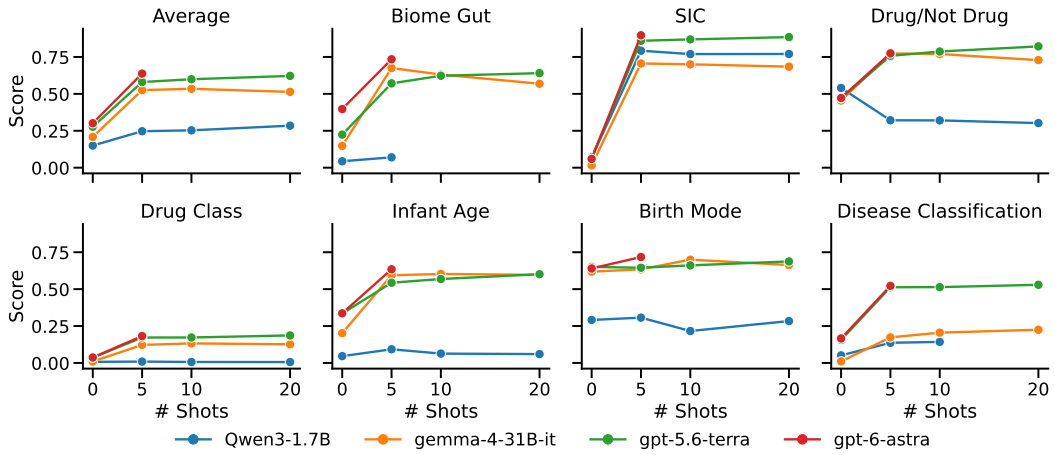


Figure 3: Few-shot classification results. 10-shot and 20-shot results are missing for some open model tasks due to them exceeding the context length. These results are missing for `GPT-6-Astra` for cost reasons. We choose to report five-shot results in the main text as they provide a reasonable trade-off between performance, coverage, and cost.

E HYPERPARAMETERS

CPT Hyperparameters:

- Hyperparamters were found by sweeping learning rates (5e-4, 1e-4, 5e-5, 1e-5, 5e-6) and observing validation loss.
- We train for 20,000 steps, with a batch size of 64 and a learning rate of 5e-5, 1000 warmup steps to a constant learning rate, a sequence length of 2048, and weight decay of 0.01. We train on 4 NVIDIA A100 GPUs.

SFT Hyperparameters:

- We found hyperparameters by sweeping learning rates (1e-4, 5e-5, 1e-4), batch sizes (16, 32, 64), and evaluation frequency (50, 100, 200) and observing val loss.
- Following (Treloar et al., 2026), we decide to stop training with early stopping on val loss with a patience of 5, evaluating every 200 steps.
- Learning rate is 1e-5, 1000 warm-up steps, constant lr. 0.01 weight decay, gradient clipping to 1.0. Batch size is 32. Loss on the user prompt is masked out.

Baseline Hyperparameters: We run hyperparameter sweeps for each of the classical machine learning baselines using the validation sets. The selected hyperparameters are in Table 7.

Multi-task models When determining mixing weights, we initially used raw per-task training token counts to weight samples from each task. However, initial experiments led to overfitting on

Table 7: Hyperparameters for classical baselines.

Dataset	Logistic Regression	Random Forest	XGBoost
Biome Gut (general)	C: 0.01	max depth: 10 n estimators: 100	max depth: 4
Biome Gut (specific)	C: 0.01	max depth: 10 n estimators: 100	max depth: 4
SIC	C: 1	max depth: None n estimators: 500	max depth: 4
Drug/Not Drug	C: 0.01	max depth: 20 n estimators: 500	max depth: 8
Drug Class	C: 0.01	max depth: 10 n estimators: 100	max depth: 4
Infant Age	C: 0.01	max depth: 10 n estimators: 100	max depth: 4
Birth Mode	C: 0.01	max depth: 20 n estimators: 500	max depth: 4
Disease Clf	C: 0.01	max depth: None n estimators: 500	max depth: 8

the smaller datasets. Instead, we use temperature sampling with temperature = 0.5, doubling the weight of Biome Gut and Disease Classification to try to avoid underfitting. Concretely, if a task t has n tokens, each sample of the task is sampled in a particular batch with probability $\left(\frac{1}{n}\right)^{0.5}$. Of the two models we tune, `Qwen3-1.7B` has been instruction-tuned while the `MicrobeText` model has not been. To alleviate differences between these two, we also mix in 10,000 samples of tulu v3 data into the SFT training data (Lambert et al., 2025). These are underweighted relative to their token count—they are longer than the classification tasks—and each one is sampled with the average probability of the other tasks rather than inversely proportional its token count. The final sample-level mixture weights are: Disease Classification: 38.3%, Biome Gut: 21.5%, Drug/Not Drug: 8.4%, SIC: 7.1%, Infant Age: 5.8%, Drug Class: 5.6%, Microbiome Reasoning: 4.4%, Synthetic Scientific Reasoning SFT: 2.4%, Tulu3: 0.9%. Note that these are sample-level weights, the knowledge-level datasets are longer, so contribute more to the loss when included. Using a patience-based early stopping on average validation loss is not effective for our multi-task models. The knowledge tasks require many steps to start to overfit due to how many tokens they have, so by the time validation loss starts to plateau, the classification tasks are overfitting. To address this, we choose a reasonable midpoint of training at 10K gradient steps.