Explanation Of Revisions PDF

The authors are grateful to the reviewers for their valuable and helpful feedback. Each comment or suggestion from the reviewers has been presented in blue Italics followed by our response and action to revise the paper accordingly

Overview of the revision

Beyond addressing reviewer feedback and enhancing overall clarity, we have also made the following significant revisions to the manuscript:

- **Data Accessibility:** We have explicitly clarified in the paper that the data for MicrobeQuest can be directly downloaded from our GitHub page: https://github.com/acl-submission/MicrobeQuest/tree/main/benchmarks. This addresses previous concerns regarding the clarity of data access.
- **Terminology Correction:** We have rectified the potentially misleading term "QA pair" and now consistently refer to our input as **"multi-modal query-response pairs"**. We appreciate the reviewers' valuable input on this important distinction.
- **Scope Refinement - Data Pipeline Details:** While several reviewers expressed interest in the technical specifics of the data extraction pipeline, task pipelining, and cross-modal alignment, this paper's core objective is to introduce MicrobeQuest as a benchmark. Presenting the intricate details of the data extraction and integration pipeline is beyond the scope of this work. Therefore, we have removed the sections that pertained to the data extraction pipeline and the data validation web system from this manuscript. We are developing a separate publication that will comprehensively cover these technical aspects.
- **Add more SOTA model for the evaluation:** We tested19 state-of-the-art (SOTA) LLM models on the MicrobeQuest benchmark. These models represent a significant portion of the SOTA LLM models currently available in the market.

Reviewer C9XS

Comment 1:

“The framework depends heavily on a pipeline that connects multiple agents, meaning if there is an error, it may cause error proprgation. It seems the error will not be caught until the validation step. Some analysis would be helpful there.”

Response:

We completely agree that the potential for error propagation within a multi-agent pipeline is a critical consideration. This is precisely why we've developed and rigorously utilized the MicrobeQuest benchmark. Our benchmark is designed not only to evaluate the overall performance of the pipeline tool, but also to specifically identify and analyze its performance across different tasks. MicrobeQuest encompasses four main tasks and eighteen sub-tasks, totaling 10,176 QA pairs specifically designed for microbiology information retrieval benchmarking (detailed in Appendix D Subtask Description). This comprehensive evaluation allows us to assess the robustness of the framework under different error conditions and refine the pipeline to minimize such propagation.

Comment 2:

Evaluation metrics is mostly still based on traditional information retrieval, even though there are many "reasoning" tasks (which can be the case the relevant materials are scattered?). I am not an expert in IR, so I am not really sure how to understand the difficulty of different tasks

Response:

Our benchmark evaluates two key reasoning abilities: Multimodal Understanding and Complex Semantic Reasoning, both of which go far beyond simple keyword matching. Multimodal tasks require models to align and integrate information across text, tables, and captions, resolving any inconsistencies. The semantic reasoning component includes seven increasingly complex tasks, such as multi-entity attribute association, multi-value priority resolution, negation parsing, logical condition reasoning, cross-paragraph entity tracking, implicit conclusion generation, and multi-instance comparative reasoning. These tasks demand deep linguistic understanding, logical inference, and specialized microbiological knowledge (examples are in Appendix D).

Despite this complexity, traditional information retrieval (IR) evaluation metrics like accuracy, F1-score, and BLEU are suitable because our tasks have objective answers. We argue that consistently correct answers across our diverse dataset of 10,176 question-answer pairs indicate effective reasoning, not just guesswork. Our approach combines complex task design with simple, interpretable evaluation: while the cognitive processes are sophisticated, we measure reasoning quality through answer correctness, ensuring scientific rigor and practical usability.

Comment 3:

More details about the data construction process can be mentioned, like in human validation, how many LLMs' wrong answers are filtered and whether humans agree with each other

Response:

Our data construction employs a two stage hierarchical validation process designed to ensure both efficiency and accuracy. The process followed by progressive human validation at different levels of expertise. During the annotation process, we implemented a robust hierarchical conflict resolution protocol to ensure consistency and accuracy:

Within-stage Conflict Handling: When disagreements occurred between two annotators within the same stage, the case was automatically escalated to the next stage for resolution. For instance, if disagreements arose during the first Stage, the case was forwarded to experts in the next Stage for final judgment.

Final Consensus Mechanism: For disputes within the highest stage (Final Stage Review), we required at least two senior experts to independently assess the case. A majority vote was employed to reach a resolution. In the rare cases where consensus could not be achieved (<2% of total cases), additional domain experts were consulted until unanimous agreement was reached.

In the manuscript, the following paragraph has been added (*please see Page 5, line 350*).

"The dataset was constructed through a two-stage annotation process involving domain experts. In the first stage, 20 microbiology students extracted predefined microbial features from the source documents. These features and associated questions were developed in collaboration with domain experts to ensure scientific relevance and to address key research needs. In the second stage, two do- main experts reviewed and refined the annotations. Each entry was independently validated by at least two experts, with disagreements resolved through discussion. In rare cases, additional experts were consulted to reach consensus."

Comment 4:

Related to the second point in weakness, some reasoning tasks have even better results than extraction, which is not very intuitive to me, as reasoning usually needs more complicated logic for the retrieval tasks. Or does that mean, some of the reasoning tasks are actually easier?

Response:

We acknowledge that the performance differences between reasoning and extraction tasks may appear counterintuitive at first glance. The performance differences reflect varying task complexities, influenced by factors like domain specificity, multi-modal requirements, and specialized knowledge, rather than just a simple extraction vs. reasoning distinction. We would like to clarify this phenomenon with the following explanations:

1. Table/Figure Extraction Tasks: These tasks involve challenges beyond simple text parsing, requiring models to process visual information and understand spatial relationships. Despite their strong text capabilities, current models struggle with visual-textual integration, explaining lower performance on these tasks.
2. Domain-Specific Extraction Tasks: Some tasks, like "Strain Culture Medium and Growth Condition Extraction," require deep microbiological knowledge and familiarity with laboratory protocols and scientific naming conventions, making them more complex.

3. Reasoning Task Performance: Certain reasoning tasks, such as "Negation and Contrast Relationship Parsing," leverage models' strong logical reasoning capabilities, which are more generalizable and well-represented in training data.

Comment 5:

The general motivation of this work is to help AI4S. But it is not clear to me how this can be used for scientific discovery. A further step may be using this tool for evaluating scientific hypotheses, instead of just answering basic questions? Since I am also not from the bio domain, some more discussion will be helpful.

Response:

A wealth of microbiology information and data remains hidden within research papers or lab reports, and researchers frequently require tools to extract and evaluate this information. The MicrobeQuest benchmark helps identify weak spots in existing models, enabling researchers to improve their methods.
In the manuscript, the following paragraph has been added (*please see Page 1, line 47*).
"*Microorganisms play vital roles in human health, agriculture, industrial biotechnology, and global ecosystems. However, experiments in microbiology often stretch over extended periods, sometimes for years. To use AI to effectively learn and acceler- ate new discoveries in microbiology, it needs data that's organized, categorized, and easily machine- readable. A vast amount of valuable microbial data is buried within existing scientific literature such as taxonomy, physiology, and cultivation conditions, but information retrieval (IR) in microbiology is exceptionally challenging, primarily due to two fac- tors: its complex multimodal nature and the sheer variety of microbiological and physicochemical attributes that characterize microbial life*"

Comment 6:

I may have missed this, but it is not clear to me how the "questions" in the tasks are generated? Are they curated by human or AI? If AI, then how they are actually picked?

Response:

The questions were developed through structured interviews with three microbiology experts, who identified key challenges and practical needs in microbial research. These insights informed the manual curation of relevant questions, ensuring the tasks' relevance and avoiding potential biases that might arise from automated question generation. We have added details about this process in Line [323-349] of the revised manuscript.

Reviewer jnTU

Comment 1:

My main concern lies in the nature and accessability of the evaluation dataset. The GitHub repository does not contains the 10,176 QA pairs described in the paper. I was not able to locate the expected answers corresponding to the queries. In the repository « MicrobeQuest/benchmarks », the files named after the specific NLP tasks (e.g. Environmental Growth Parameter Extraction) do not contain the actual answers but only the prompts and a small number of demonstration examples that are duplicated may times for each input document ID. Moreover, it appears that the document corpus of 127 documents cannot be reconstructed as the references are no provided - some of which are not open access — and the retrieval process is not described. Calling this resource « benchmark » in both the paper and the repository is therefore quite misleading.

Response:

We understand the reviewer's concern about the need for clearer clarification of the accessibility and structure of the dataset. The 10,176 QA pairs mentioned in the paper are indeed included, but we recognize that the current structure of the GitHub repository may cause some confusion. To address this, we have updated the GitHub Readme to guide users to the appropriate data file path. All data is located in the Github Page(https://github.com/acl-submission/MicrobeQuest/tree/main/benchmarks) directory, which contains 18 JSON files, each corresponding to a specific NLP task. Each task file includes test questions, a chain-of-thought (COT) prompting section, few-shot examples for model priming, explicit annotations outlining the expected answer type, and the corresponding ground truth answers that serve as the benchmark for evaluation.

Regarding the document corpus, due to copyright restrictions imposed by academic publishers, we are unable to provide the complete set of 127 original documents. However, to support evaluation and ensure reproducibility, we have provided DOI links to all referenced publications and included automated Python scripts in our GitHub page to help retrieve the corresponding PDF articles. While the full text cannot be distributed due to legal constraints, we believe this still offers a representative and standardized evaluation framework. We are committed to helping researchers reconstruct and evaluate the dataset as fully as possible within these legal and ethical boundaries.

Comment 2:

A second concern relates to the few-shot training example quality, as some of them appear to be incorrect. For example, in the QA pair "question": "What is the source environment for this microbe?", "text": "The strain thrives in deep subsurface shale formations where it contributes to biogeochemical cycles.", "expected_answer":

"[object Object]Subsurface[object Object]" the term "subsurface" is treated as the source environment, which it is not since the expected answer is either the whole term "deep subsurface shale formations" or "sedimentary rock"

Response:

We acknowledge that in the cited example—where the QA pair extracts "subsurface" from the sentence "The strain thrives in deep subsurface shale formations where it contributes to biogeochemical cycles"—the extracted answer may appear overly general. However, this is a direct result of the current system design, which maps terms to a controlled environmental ontology.

In this ontology, specific composite terms such as "deep subsurface shale formations" or "shale formation" are not available as distinct entries. Therefore, the extraction process defaults to the most specific matching concept, in this case "subsurface." While "sedimentary rock" could be considered a more accurate environmental descriptor, it is not explicitly stated in the text and thus not inferable under our current pipeline without external reasoning.

We recognize the importance of high-quality training examples and agree that improvements are needed. As part of our ongoing work, we are actively exploring ways to refine the mapping granularity in the structured formats in the Environmental Ontology (EnvO), and improve example filtering and validation to better align with domain-specific expectations.

Comment 3:

Another example is the following QA pair: "question": "What is the full name of a strain (species+strain)?", "text": "The medium was bubbled with nitrogen (N$_2$) for 10 minutes to create an anaerobic environment. The culture used Methanobacterium ivanovii strain Ivanov for methanogenesis studies.", "expected_answer": "[object Object] Methanobacterium ivanovii Ivanov [object Object]" However, Methanobacterium ivanovii Ivanov does not appear in the text and does not correspond to any valid taxonomic entry. The correct full name should be Methanobacterium ivanovii Jain et al. 1988. These incorrect examples are repeated dozens of times throughout the dataset, raising concerns about the quality and reliability of the few-shot demonstrations used in the benchmark to instruct the LLMs.

Response:

In the cited example, the text contains the phrase "Methanobacterium ivanovii strain Ivanov", from which our system extracts the species name "Methanobacterium ivanovii" and the strain designation "Ivanov". This information corresponds to entries in public databases such as LPSN/BacDive (e.g.,https://bacdive.dsmz.de/strain/168848). We acknowledge that this format differs from the full formal scientific name, "Methanobacterium ivanovii Jain et al. 1988." However, our current system is designed to extract and normalize names based solely on textual evidence, without cross-validating against external taxonomic registries.

Comment 4:

The NLP tasks should be better defined including the task objective, the microbial information, and the output formal representation (attributes, types). External references when used should also be given. Listing the names of the task is not sufficient. Moreover, some task names are misleading. For instance, the examples in the gtihub/benchmark data of the so-called « strain entity recognition and normalization » task does not include normalization and the reference taxonomy is not provided (NCBI taxonomy ? LPSN ?). The paper also lacks details on the way the tasks are pipelined, for example, how the habitats extracted by the Strain Attribute Semantic Categorization task are linked to the given strain, or how the Cross-modal Alignment Agent is able to link the data extracted from different sources by previous tasks. These are well-know as very challenging tasks

Response:

We understand the reviewer's concern about the need for clearer clarification of the task definitions and output representations. We acknowledge that the original manuscript only provided a brief overview of the tasks. To address this, we have provided detailed task definitions in Appendix D, which includes task objectives, input microbial information, output representations, evaluation criteria, and specific examples. In the revised version, we will enhance the main manuscript by adding accurate references to Appendix D Line [319-322] guiding the reader to refer to the appendix for the full task definitions.

Clarification of the "Strain Entity Recognition and Normalization" Task:

We understand the concern regarding the task name being misleading. The "normalization" in this task refers to the standardization of strain name representations for consistent identification. Specifically:

Recognition: Identifying strain mentions in various formats within the text.

Normalization: Standardizing mentions to a complete "species + strain" format (e.g., "E. coli K-12" to "Escherichia coli K-12").

We will clarify the specific meaning of this task in Appendix C Lines [834–901] of the revised manuscript

And we understand your interest in the technical details of task pipelining and cross-modal alignment. However, as this paper primarily focuses on presenting MicrobeQuest as a benchmark, the detailed technical implementation of the data extraction and integration pipeline is beyond the scope of this work. We are preparing a separate manuscript that will focus specifically on these technical aspects. In the meantime, we will remove pipeline methodology from this paper

Comment 5:

Given the incorrect demonstration examples in the prompts, the lack of task specification, and the absence of the ground truth dataset (i.e. the 10,176 QA pairs) used for the evaluation, it is unclear how the models were evaluated and how such high performance could be achieved on these complex tasks

Response:

We apologize for any confusion regarding the dataset availability and evaluation process. We want to clarify that the complete 10,176 QA pairs with ground truth answers are indeed available in the GitHub repository, located in the https://github.com/acl-submission/MicrobeQuest/tree/main/benchmarks directory. Each task file includes the corresponding ground truth answers. While we cannot redistribute the full-text documents due to copyright constraints, the ground truth answers are based on our original annotations and are freely available. We also provide automated evaluation scripts, and the process is fully reproducible using the provided data.

Comment 6:

The evaluation results should be compared to state-of-the-art performances obtained on similar tasks. The Bacteria Biotope shared task involves normalization and the extraction of entities and relation of similar information as MicrobeQuest (phenotypes, habitats and taxon including strains). This benchmark should be cited and used as a point of comparison of SOTA methods with MicrobeQuest's method performances. The author claims about the novelty of their work relies on an inaccurate description of related work. In lines 072 to 080 prior approaches are reduced to « Traditional IR systems like BM25 (Robertson et al., 2009) or TF- 073 IDF (Ramos et al., 2003) which overlook more advanced and fine-grained NLP methods. The MicrobeQuest methods belongs to Information extraction or QA more than IR. Such methods have already been evaluated in the Bacteria Biotope task, in a manner comparable to the objectives and structure of MicrobeQuest.

Response:

We sincerely appreciate the reviewer's valuable feedback regarding the comparison with the Bacteria Biotope (BB) shared task. We acknowledge that the BB shared task has made significant contributions to the field of biomedical NLP and serves as an important benchmark for microbiology-related information extraction. We agree that a more comprehensive discussion of the BB task in the related work section would enhance the completeness of our paper.

However, we would like to further clarify that our work is complementary to the BB shared task in terms of objectives, as both focus on different dimensions of microbiology knowledge extraction and serve different research goals and practical applications.

Task Scope and Complexity: The BB shared task focuses on three core information extraction tasks: entity normalization (BB-norm), relation extraction (BB-rel), and knowledge base construction (BB-kb). These tasks are well-defined and provide structured evaluation frameworks for traditional NLP approaches. In contrast, MicrobeQuest encompasses 18 different NLP subtasks that were specifically designed in collaboration with microbiology domain experts to address practical information needs encountered in real research scenarios. Our broader task coverage reflects the diverse and complex information requirements that researchers face when consulting microbiology literature.

Methodological Comparison: The BB shared task has primarily evaluated traditional machine learning methods, including Conditional Random Fields (CRFs) for named entity recognition, Support Vector Machines (SVMs) for relation extraction, and neural network models based on LSTMs and CNNs. While recent iterations (since around 2019) have also incorporated pre-trained language models like BERT, the BB evaluation framework was originally designed for earlier NLP technologies. In contrast, the core objective of MicrobeQuest is to evaluate the performance of state-of-the-art large language models (LLMs) such as GPT and Claude, which represent the latest developments in NLP.

Evaluation Paradigms: The BB task employs traditional information extraction evaluation paradigms, using metrics such as precision, recall, and F1-scores, with emphasis on entity boundary matching accuracy and relation triplet extraction. This evaluation approach is well-suited for assessing structured knowledge extraction systems. MicrobeQuest adopts a question-answering evaluation paradigm that, while using traditional information extraction evaluation metrics, directly measures models' ability to understand and answer specific microbiology questions. This approach is more aligned with researchers' actual behavior when reading scientific literature—they typically search for answers to specific questions rather than extracting predefined entity and relation types.

Data Characteristics and Knowledge Representation: The BB dataset consists of 215 PubMed abstracts and 177 full-text excerpts, requiring complex entity and relation annotations within a structured framework. The task emphasizes standardization using the NCBI taxonomy system and OntoBiotope ontology, with fixed entity and relation types. MicrobeQuest involves 127 papers and constructs 10,176 question-answer pairs, covering a broader range of microbiology topics. We employ a flexible question-answering knowledge acquisition approach that does not rely on specific ontologies, supporting diverse question formats and natural language interactions.

We view our work as complementary to, rather than competitive with, the BB shared task. The BB task provides valuable standardized evaluation for structured knowledge extraction and has established important foundations for microbiology NLP research. MicrobeQuest offers unique contributions by evaluating modern large language models' capabilities in microbiology, providing broader task coverage, and employing question-answering evaluation that better reflects real-world application scenarios.

Comment 7:

The general objectives and processes of MicrobeQuest should also be compared to the Omnicrobe application which is inaccurately characterized in the paper as relying on "word statistics.". A thorough comparison is needed in terms of data sources, methods, scope and intended objectives. Omnicrobe's sources also include microbe collections, genetic databases and papers. The phenotypes and habitats categories used in Omnicrobe closely resemble those in MicrobeQuest (although their *labels may differ). Omnicrobe includes fine-grained normalization of microbial phenotypes (morphology, physiology, ..), habitats (isolation, growth medium) and taxa all structured by relations ships and by the OntoBiotope classes and NCBI taxa.*

Response:

We acknowledge that our initial characterization was oversimplified, and we apologize for the inaccuracy. Upon further review, we recognize that Omnicrobe employs more sophisticated methods, including ontology-based and NLP-driven text mining technologies, to address the limitations inherent in traditional approaches. Omnicrobe integrates multi-source data and utilizes structured ontologies such as OntoBiotope and NCBI taxonomy to categorize microbial habitats, phenotypes, and taxa. This goes beyond simple keyword matching, allowing Omnicrobe to handle complex, multi-faceted microbiological data more effectively than traditional IR systems.

In contrast, MicrobeQuest is an evaluation benchmark designed to provide standardized assessment metrics for information retrieval (IR) models applied to microbiology tasks. Unlike Omnicrobe, which focuses on comprehensive data integration and knowledge discovery, MicrobeQuest provides a framework to evaluate the accuracy and effectiveness of models in microbial data extraction and reasoning.

We have updated our manuscript to more accurately reflect Omnicrobe's advanced methodologies. In the manuscript, the following paragraph has been added (*please see Page 1, line 71*).

"In contrast, Omnicrob employs ontology- based NLP and text mining techniques to manage the complex relationships between microbial habi- tats, phenotypes, and uses (Dérozier et al., 2023)."

We deeply appreciate the reviewer's input, which has enabled us to position our work more precisely within the broader research landscape and give proper credit to Omnicrobe's technical achievements

Comment 8:

The selection of the source papers and databases of MicrobeQuest should be explained. The project appears to focus primarily on environmental microbes, yet includes sources that raise questions. For example, the journal Standards in Genomic Science changed name in 2018 and , and Antonie van Leeuwenhoek is not considered a leading journal in the field. Additionally, MIRRI-ERIC could be a more relevant microbial resource than NCIMB, and databases like GOLD or JCM might be more appropriate choices for comprehensive microbial data

Response:

1. Scope of MicrobeQuest and Source Papers:

While MicrobeQuest places a strong emphasis on environmental microbes, this reflects the broader trend in microbial taxonomy and strain characterization: the majority of publicly available strains with rich phenotypic and ecological metadata have been isolated and described from environmental sources. Our selection of strains was guided by data availability and representativeness, rather than by a focus exclusively on environmental origins. Importantly, our dataset also includes a diverse range of non-environmental microbes. Specifically, approximately 6% of the strains are derived from the gut, and 10% are associated with industrial contexts. These strains complement the ecological breadth of the environmental isolates and contribute to a more comprehensive microbial trait landscape.

Regarding the inclusion of journals such as Standards in Genomic Science and Antonie van Leeuwenhoek, we acknowledge that these may not be the most high-impact venues. However, they have consistently published standardized strain descriptions, often with curated genome-linked metadata and structured phenotype fields that are well-suited for extraction. We will clarify the rationale for including these sources and note the name change and discontinuation of Standards in Genomic Science where appropriate in the revised manuscript.

2. Database Selection:

Our microbial medium dataset has been constructed using information from multiple major microbial resource centers (DSMZ, ATCC, and NCIMB). These repositories were selected based on the richness of their structured medium formulations, global recognition, and historical contributions to microbial systematics. Each of these culture collections provides: (1)Explicit linkage between strains and their growth media; (2)Well-documented medium compositions with standardized component naming; (3)Reliable strain metadata, often including ecological and physiological characteristics.

We recognize that the reviewer points out other highly relevant databases such as MIRRI-ERIC, JCM, and GOLD. These resources are indeed authoritative and complementary. In fact, there is substantial overlap between the types of data provided by these and the collections we have already incorporated. For example:

(1)Like DSMZ and ATCC, JCM provides detailed media formulations and growth conditions, and we are actively integrating JCM data into MicrobeQuest.

(2)MIRRI-ERIC, as a European integrative infrastructure, aggregates information from several national collections, including data similar to those from NCIMB and DSMZ.

(3)GOLD offers genome-centric metadata and environment annotations, which align well with the structured ecological descriptors used in MicrobeQuest. We are currently assessing GOLD integration.

We will revise the manuscript to explain our data source strategy more clearly, emphasize the inclusion of clinical and industrial strains alongside environmental microbes, and explicitly discuss the similarities and complementarities between DSMZ/ATCC/NCIMB and resources such as MIRRI-ERIC, JCM, and GOLD. We will also outline our ongoing efforts to incorporate these additional resources in future updates of MicrobeQuest.

Comment 9:

Line 892 : The task described in the prompt template aims to link different descriptions of the same strain. The paper should clarify how these descriptions are represented in the input question and how the output format is structured to distinguish between linked and non-linked strain entities

Response:

We understand the reviewer's concern about the need for clearer explanation of input question representation and output format structure in our strain entity linking task.
Clarification of Task Design:

1. Task Context: We address the complex scenario of single publications containing multiple strain entities. In microbiological literature, the same strain often appears under different names or identifiers (e.g., laboratory codes, culture collection numbers, type strain designations), making the identification of whether different names refer to the same strain entity a significant challenge.
2. Input Question Format: Our input questions adopt a direct comparison format, such as “Is strain SL43 the same as JCM 11886?”, explicitly asking whether two strain identifiers refer to the same entity.
3. Input Text Representation: The input text contains relevant information needed for entity linking decisions. For example, text passages containing expressions like “strain KIN24-T80T = DSM 22612T = JCM16315T” clearly indicate equivalence relationships between different identifiers.
4. Output Format Structure: The model outputs explicit judgments (e.g., "True" or "False") based on textual evidence. For linked entities, the model must identify linguistic markers indicating equivalence relationships (such as "=" symbols, parenthetical notation); for non-linked entities, the model must distinguish their different characteristics.

We have provided technical details of the task’s input/output formats and examples in Figure 2 and Appendix D of the paper.

Comment 10:

Line 909 : the model is given a list of example key physiological characteristics of strains. However, morphology is not a physiological trait. Also, it is unclear what is meant by "strain type" — does this refer to the strain type for species? The paper should explain how the extracted results are intended to populate knowledge bases given that the model is not provided with a comprehensive list of expected characteristic types with their labels.

Response:

We understand the reviewer's concern about the need for clearer clarification of the terminology and design choices in our task definition for physiological trait extraction

1. Morphology vs. Physiology: You are correct in noting that morphology, strictly speaking, is not a physiological trait. In our task definition, we adopted a broader

operational interpretation of "physiological characteristics" to include commonly used microbial diagnostic features — such as cell shape, motility, oxygen tolerance, and Gram reaction — which are often reported together in strain descriptions and used in microbial classification and identification. While “morphology” is typically categorized under structural or taxonomic features, it was included due to its close association with phenotypic trait profiles in microbial strain literature. We will revise the text to clarify this broader usage and distinguish between strictly physiological traits and morphological descriptors.

2. Clarification of “Strain Type”: The term “strain type” was intended to refer to designations such as “type strain” or “reference strain”, which often carry taxonomic or functional significance in microbial databases. However, we acknowledge that the wording is ambiguous and will revise the manuscript to clearly define what is meant by this term and how such labels are parsed from input text.
3. Trait Schema and Knowledge Base Integration: Regarding the reviewer's comment that the model does not provide a comprehensive list of expected traits and labels, we would like to clarify that the required output format is explicitly outlined in the "Note" section of our prompt template. This section guides the model's feature extraction process, ensuring that the model understands the expected traits and response format. In the revised version, we will further elaborate on our COT guidance in the Appendix C to clarify the model's task handling logic

Comment 11:

Line 923 : The prompt instructs the LLM to extract and categorize the identified strain attributes according to established taxonomic and semantic frameworks. However, can it be done since the « established taxonomic and semantic frameworks » are not given. These frameworks are not defined or provided. It is unclear what the expected output format is, and how the extracted results can be reliably aggregated if the categories are freely generated by the LLMs without guidance or constraints

Response:

We understand the reviewer's concern regarding the absence of a clear taxonomic and semantic framework. To clarify, the LLM was provided with predefined taxonomic and semantic frameworks as part of our experimental setup. These frameworks, which are defined in the Note section of the prompt template, outline specific categories and classification schemes for strain attribute extraction. The LLM operates within these predefined boundaries, ensuring consistent and reliable output without arbitrary category generation. We have updated the manuscript to explicitly mention these framework specifications at Appendix C Line[820-1014]

Comment 12:

Line 272-273 : PubMed is not an ontology.

Response:

We appreciate the reviewer's comment regarding PubMed. PubMed is a literature database, not an ontology. In our study, we utilized the MeSH (Medical Subject Headings) ontology, a controlled vocabulary system within PubMed, to standardize search terms for comprehensive retrieval of microbial studies.

Comment 13:

Line 477-478 : translation is not involved. Figure 3 is not easily readable. Fgure 4 should include the number of documents and the publication date interval.

Response:

We thank the reviewer for their valuable feedback on Figure 3 and Figure 4. Based on the suggestions:
Figure 3 has been redrawn with an improved layout and color scheme, and the caption now includes a brief explanation for better clarity.
Figure 4 now includes the number of documents associated with each database, and the caption clearly mentions the publication date range.
Additionally, we have revised Lines 432 to address the confusion regarding the term "translation" These changes aim to improve the clarity and completeness of the manuscript, and we sincerely appreciate the reviewer's input.

Comment 14:

I recommend either targeting a venue that allows a more comprehensive presentation, or focusing this submission on a core, original component of the corpus building pipeline that can be fully developed and evaluated within the available space. This could strengthen the clarity and impact of the contribution.

Response:

Thank you for your thoughtful and constructive feedback. We are glad that our revisions addressed your concerns and appreciate your kind suggestion to consider submitting our work elsewhere. However, we firmly believe that EMNLP is the most appropriate venue for this submission. Our paper introduces MicrobeQuest, the benchmark specifically designed for information retrieval in microbiology, featuring 10,176 expert-validated QA pairs across 18 domain-specific tasks. EMNLP's Theme Track on Interdisciplinary Recontextualization and Benchmark Track closely align with the goals of our work. As EMNLP emphasizes, "the core interests of the ACL community are rooted in human-language technologies but also have broad reach into other fields." Our benchmark embodies this interdisciplinary spirit by bridging NLP and microbiology—a field where vast amounts of critical knowledge remain locked in unstructured literature. Researchers often need tools to extract, organize, and evaluate microbiological information hidden in scientific papers and lab reports. MicrobeQuest serves as both a diagnostic tool for identifying the limitations of current IR models and a foundation for developing improved, domain-adapted methods. In future versions of the manuscript, we will further elaborate on how our benchmark can support hypothesis evaluation and enable more advanced scientific exploration. We believe EMNLP provides the right audience and context to foster the interdisciplinary dialogue our work is meant to inspire. Thank you again for your review and support.

Reviewer V4Z7

Comment 1:

More insights into the results would be helpful. Currently, the analysis primarily focuses on surface-level observations, such as which model performs better on which task. Including deeper insights and examples of cases where models commonly perform poorly could enhance the impact of the paper and offer valuable directions for future research aimed at improving LLMs.

Response:

Response: The manuscript has been revised based on your comments. To enhance understanding of our model's performance and error cases, we have added two appendices:

Appendix F: This appendix provides a concise overview of the model's performance, including accuracy, F1, and BLEU scores across various benchmark tables and charts. It aims to present a comprehensive evaluation of the model's strengths, weaknesses, and applicability across different tasks and metrics.

Appendix G: This section, titled "Error Analysis," categorizes the key error types identified during our tasks and provides illustrative examples in accompanying tables, offering insights into areas for improvement.

Meta-review Area Chair fsay

Comment 1:

A better description of the construction of the IR benchmark, e.g., data construction, human annotation and agreement, task objectives, inputs and outputs, example selection for few-shot learning. In addition, evaluation measures should be task-dependent and not uniform across tasks.

Response:

We've updated the manuscript based on the reviewers' suggestions. We've included a more detailed description of the IR benchmark's construction and revised Appendix C, which provides input/output examples. As for evaluation measures, we have applied three different measures : F1 score, accuracy, and the Bilingual Evaluation Understudy (BLEU) score. More information can be found at Line[411-416]

Comment 2:

Better motivation on why using only QA is appropriate for evaluating the extraction, retrieval, and reasoning tasks in the biology field.

Response:

In this revised paper, we've clarified that we're using multi-modal query-response pairs as input, not QA (question-answer) pairs. We appreciate the reviewer's feedback, as "QA pair" was indeed misleading.

Comment 3:

Task evaluation is minimal. Present detailed results by task group and give indications on the analysis of successes and failures by task group as well. The comparison with the SOTA model should also be presented.

Response:

Figure 3 illustrates the performance of 19 state-of-the-art (SOTA) LLM models on the MicrobeQuest benchmark. These models represent a significant portion of the SOTA LLM models currently available in the market. The results for the top six performing models are specifically detailed in Table 3, while the performance of the remaining models is presented in Appendix F. Furthermore, we provide an in-depth error analysis in Appendix G.