MicrobeQuest: A Multimodal Benchmark for Information Retrieval in Microbiology

Anonymous ACL submission

Abstract

AI for Science (AI4S) is reshaping research paradigms across numerous disciplines. In microbiology, multimodal data (text, images, tables, and charts) exist in scientific literature and public databases to understand the complex relationship between microbial strains and their environment. However, current benchmarks are either general-purpose or designed for disciplines such as material or biomedical sciences, lacking one specific for microbial sciences. Here, we developed MicrobeQuest, the first comprehensive, multimodal benchmark with 10,176 Question-Answer (QA) pairs for microbiology-specific information retrieval to take advantage of the vast amount of available information in microbiology. We first developed an expert-in-the-loop platform (MicrobeCollect) to acquire and annotate microbiological data. We then demonstrated its utility by benchmarking 17 state-of-the-art (SOTA) information retrieval (IR) methods. This yielded crucial performance insights and established a robust foundation for future IR advancements in microbiology. All benchmark resources, including code and datasets, are publicly available at https://github.com/acl-submission/MicrobeQuest.

1 Introduction

The rapid advancement of AI for Science (AI4S) is fundamentally transforming research capabilities and methodologies across scientific disciplines, enabling unprecedented discoveries and innovations. The biological sciences in general have greatly benefited from recent developments in natural language processing (NLP), with notable tools like Evo and AlphaFold advancing our understanding of genomics and protein structures (Nguyen et al., 2024; Abramson et al., 2024). Increasingly, AI is demonstrating strong potential in analyzing and predicting relationships in large-scale and complex biological datasets (Reiman et al., 2017; Hackmann and Zhang, 2021; Hoarfrost et al., 2022; Koblitz et al., 2024), underscoring its growing role in accelerating biological discovery.

Microorganisms play vital roles in human health, agriculture, industrial biotechnology, and global ecosystems. They are ubiquitous and highly diverse, yet the scientific knowledge about microbial traits, such as taxonomy, physiology, and cultivation conditions, remains fragmented across research articles and heterogeneous databases. Despite centuries of microbiological research, we still lack effective tools to systematically extract, annotate, and integrate microbial data at scale. This limitation impedes efforts to connect phenotypic traits with emerging insights from AI-driven genomics and proteomics, thereby hindering progress in understanding microbial function. Addressing this gap is essential for unlocking knowledge hidden in existing literature and enabling AI systems to support discovery in microbiology and its applications.

Microbiology data presents significant challenges for information retrieval (IR), primarily due to two factors: its complex multimodal nature and the sheer variety of microbiological and physicochemical attributes that characterize microbial life. A previous study used species taxonomy to integrate microbial phenotypic data from multiple databases, but this effort depends on well-curated databases (Madin et al., 2020). Traditional IR systems like BM25 (Robertson et al., 2009) or TF-IDF (Ramos et al., 2003), which rely on keyword statistics, have been applied to extract information from literature abstracts (Zafeiropoulos et al., 2022; Dérozier et al., 2023), but are insufficient for capturing the nuanced microbiological information embedded in unstructured and semantically rich full texts. The most comprehensive structured resources to date are produced from years of manual curation of culture collections, yet these remain labor intensive and still missing the more comprehensive knowledge from the vast primary literature

(Oberhardt et al., 2015; Schober et al., 2025).

Advanced IR systems that integrate Large Language Models (LLMs), multimodal LLM (MLLM) architectures and agentic systems offer a promising opportunity to improve microbiological IR tasks and reasoning from both primary literature and various microbial databases. These advanced IR systems can process complex queries and heterogeneous data types, enabling more accurate extraction and deeper semantic understanding of microbiological content. However, existing IR system benchmarks are designed for chemical and biomedical purposes (Thakur et al., 2021; Edwards et al., 2021; Gupta et al., 2024), and do not reflect the specific challenges in microbiology. A reproducible and microbiology-specific benchmark would provide a robust foundation for evaluating advanced IR systems and accelerating the development of AI tools for microbiology discoveries.

In this work, we introduce two integrated systems that advance IR systems for multimodal microbiological data: a collection platform and a benchmark. These tools are designed to improve information access in this discipline and establish standardized evaluation practices that can benefit the microbiology and AI4S communities. Our main contributions are:

1. **MicrobeCollect**: An online expert-in-the-loop IR platform that integrates multimodal input, agentic retrieval, and human expert verification to significantly accelerate microbiological data acquisition. This system bridges the gap between automated processing and domain expertise.
2. **MicrobeQuest**: A comprehensive, fine-grained benchmark for microbiology-specific IR tasks, constructed using our MicrobeCollect platform in collaboration with domain experts. Extract microbiology-specific data from journals and microbial culture collection databases. It fills a critical evaluation gap and serves as the first reproducible benchmark specifically designed for microbiology-focused IR research.
3. **IR Model Performance Evaluation**: We conduct an extensive evaluation of 17 SOTA IR models on the MicrobeQuest benchmark. This analysis provides performance comparisons across models in the microbiological domain, offering actionable insights to guide future research at the intersection of NLP and microbiology.

2 Related Work

2.1 Information Retrieval (IR) Methods

The exponential growth of diverse data in microbiology—textual documents, figure images, and summary tables and charts—demands efficient and accurate information retrieval, but existing methods face persistent challenges. Traditional pipeline tool IR methods, using techniques like TF-IDF (Robertson et al., 2009), BM25 (Ramos et al., 2003) and NLP (Zafeiropoulos et al., 2022; Dérozier et al., 2023), offer efficiency but lack deep semantic comprehension vital for complex scientific queries (Zhang et al., 2025). These pipeline approaches often remain insufficient for nuanced conceptual understanding, especially in multimodal contexts common in microbiology. Latest IR systems incorporating LLMs represent a significant advancement in semantic understanding (Brown et al., 2020), but are prone to factual inaccuracies and hallucinations without robust grounding (Bang et al., 2023). They often fail to ensure the deep, verified domain-specific knowledge fidelity crucial for scientific research in fields like microbiology. LLM mitigation techniques like fine-tuning (Lu et al., 2025), RAG (Siriwardhana et al., 2023) , or prompting (Giray, 2023) have demonstrated promising improvements across various domains, including legal (Cui et al., 2023), finance (Wu et al., 2023) and medicine (Thirunavukarasu et al., 2023). Extending this, MLLMs like GPT-4V (Yang et al., 2023) and Deepseek-vl (Lu et al., 2024) can handle diverse data type retrieval tasks but still struggle significantly with processing long scientific documents, and performing the fine-grained reasoning needed for scientific research (Zong et al., 2024). The agentic contextual retrieval paradigm offers further advancement, employing autonomous agents for dynamic retrieval strategies, planning, and tool use (Zhang et al., 2025). However, implementation complexity and the demands of reasoning with domain-specific knowledge bases limit their impact in specialized domains (Singh et al., 2025).

To address challenges, we propose MicrobeCollect: a multi-agentic IR system deeply integrated with a dedicated domain knowledge base and expert-in-the-loop. This system aims to synergize the dynamic capabilities of AI agents with the ac-

Category	Subtask	TREC	MS MARCO	MMLU-Pro	BEIR	SciAssess	μ-Bench	CBLUE	SciRIFF	MicrobeQuest
Microbial Domain	Microbial Domain Included	×	×	×	×	×	×	×	×	✓
Structured Information Extraction	Strain Entity Recognition and Normalization	✓	✓	×	×	✓	×	✓	✓	✓
	Strain Entity Resolution	×	×	×	×	×	×	×	×	✓
	Strain Taxonomy Extraction	×	×	×	×	×	×	×	×	✓
	Strain Physiological Characteristic Extraction	×	×	×	×	✓	×	×	×	✓
	Environmental Growth Parameter Extraction	×	×	×	×	✓	✓	×	×	✓
	Strain Attribute Semantic Categorization	×	✓	×	×	×	✓	✓	✓	✓
	Strain Culture Medium and Growth Condition Extraction	×	×	×	×	×	×	×	×	✓
Multimodal Understanding	Table-based Strain Attribute Extraction	×	×	×	×	✓	×	×	✓	✓
	Figure-based Strain Attribute Extraction	×	×	×	×	✓	✓	×	×	✓
	Multimodal Strain Attribute Reasoning	×	×	×	×	✓	×	×	×	✓
Complex Semantic Reasoning	Multi-Entity Attribute Association	×	✓	×	×	✓	×	✓	✓	✓
	Multi-value Priority Resolution	×	✓	×	×	×	×	×	×	✓
	Negation and Contrast Relationship Parsing	×	×	×	×	×	×	×	×	✓
	Logical Condition Reasoning	✓	✓	✓	×	✓	✓	×	×	✓
	Cross-Paragraph Entity Tracking	×	✓	×	×	×	×	×	×	✓
	Implicit Conclusion Generation	×	✓	✓	×	×	×	✓	✓	✓
	Multi-Instance Comparative Reasoning	×	×	×	×	×	×	×	×	✓
Layout Structure and Semantic Region Recognition	Semantic Document Region Extraction	×	×	×	×	×	×	×	×	✓
Task Paradigm	INPUT	Text	Text	Text	Text	Image(Table) & Text	Image	Text	Text	Image(Table& Chart) & Text
	OUTPUT	Text	Text	Text	Text	Text	Text	Text	Text	Text

Table 1: A comparison between MicrobeQuest and existing IR benchmarks, including TREC (Voorhees et al., 2005), MS MARCO (Bajaj et al., 2018),MMLU-Pro (Wang et al., 2024), BEIR (Thakur et al., 2021), SciAssess (Cai et al., 2024), μ-Bench (Lozano et al., 2024), CBLUE (Zhang et al., 2022), SciRIFF (Wadden et al., 2024). The presence of a specific task within a benchmark is indicated by ✓; its absence is indicated by ×.

curacy and reliability of expert-curated, structured microbiological knowledge.

2.2 Benchmarks for Information Retrieval

General-purpose benchmarks play a fundamental role in advancing IR systems, offering standardized frameworks for evaluating effectiveness, objectively comparing techniques, and identifying strengths and weaknesses. The impact of benchmarking is evident in influential evaluations: long-standing TREC collections have offered diverse tasks over decades (Voorhees et al., 2005); large-scale datasets such as MS MARCO have spurred advancements in passage ranking and question answering (Bajaj et al., 2018). The recently introduced BEIR suite aggregates 18 publicly available datasets to evaluate zero-shot generalization in diverse text retrieval tasks and domains (Thakur et al., 2021). Beyond traditional IR tasks, the evaluation landscape increasingly includes benchmarks assessing core capabilities of large language models (LLMs) relevant to modern IR systems. MMLU-Pro, derived from academic exams and textbooks across 14 diverse domains (Wang et al., 2024); BIG-bench, focusing on tasks designed to be beyond current language model capabilities (Srivastava et al., 2023); HaluEval, specifically designed to evaluate hallucination tendencies in LLMs (Li et al., 2023).

While general-purpose benchmarks effectively evaluate the overall capabilities of IR systems and important for advancing this field, they often lack focus on highly specialized domains or AI4S topics that demand specific IR strategies and specialized knowledge bases (Cai et al., 2024). Recognizing this, specialized IR benchmarks have been developed for those specialized areas like biomedical, chemical or material sciences. AI4S benchmarks such as SciAssess focus on four different domain-specific requirements, like extracting complex chemical entities or disease relationships, providing a more targeted evaluation than general-purpose benchmarks (Cai et al., 2024). μ-Bench is dedicated to assessing large language models' perceptual and cognitive capabilities in analyzing biological and pathological microscopy images (Lozano et al., 2024); CBLUE, as a Chinese biomedical language understanding evaluation platform, encompasses entity recognition, relation extraction, and text classification tasks in medical texts (Zhang et al., 2022); SciRIFF primarily concentrates on information extraction and content summarization in the biomedical domain (Wadden et al., 2024).

As Table 1 shows, existing general-purpose and dedicated AI4S IR benchmarks inadequately ad-

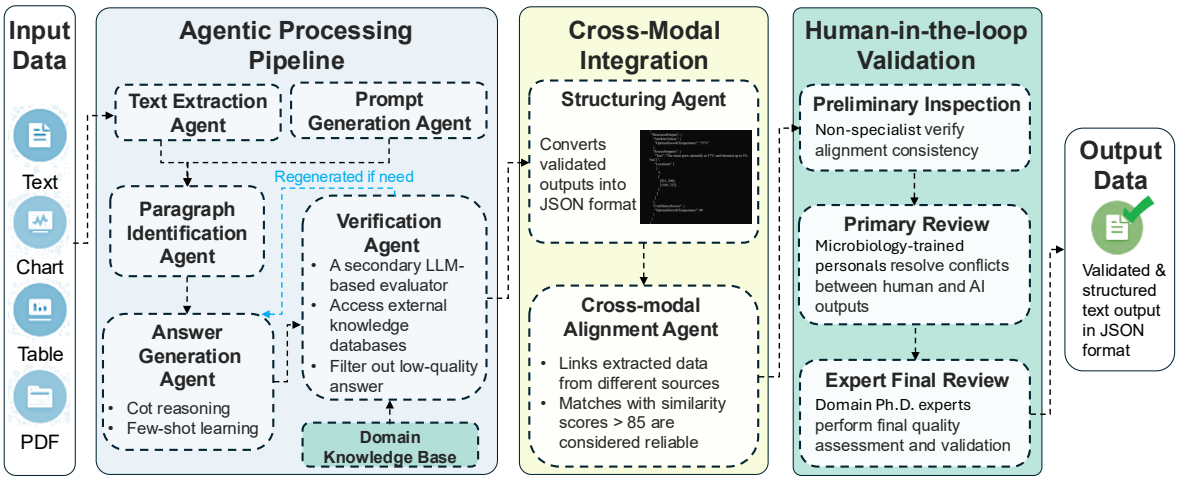


Figure 1: The MicrobeCollect Data Retrieval and Annotation Workflow involves an agentic processing pipeline for information retrieval, cross-modal integration to standardize data structures and validate against external databases, and a three-step Human-in-the-Loop process to ensure output data accuracy.

dress microbiology's unique complexities, including specialized strain terminology, diverse data integration, distinct experimental contexts, and specific structured information extraction (e.g., strain taxonomy, physiological characteristics, and cultivation conditions). This critical gap hinders the development and evaluation of tailored IR systems for this field. Addressing this gap, we introduce MicrobeQuest, which, to the best of our knowledge, is the first benchmark specifically designed for evaluating diverse information retrieval capabilities on the subject of microbiology.

3 Benchmark Construction

3.1 MicrobeCollect Workflow

MicrobeCollect was designed to accurately prepare large amounts of annotated data for IR systems benchmarks and AI applications. As shown in Figure 1, this agentic IR framework utilizes specialized agents in a complete pipeline for extracting and verifying structured data from multimodal scientific literature. The first **Text Extraction Agent**, which uses Mistral OCR to convert PDFs into structured multimodal Markdown. Compared to traditional OCR systems (e.g., Tesseract (Kay, 2007)) and Transformer-based models (e.g., Donut (Kim et al., 2022)), Mistral offers improved recognition of LaTeX formulas and complex table layouts commonly found in academic documents. The resulting content is processed by the **Prompt Generation Agent**, which constructs adaptive, domain-aware prompts. Expert-defined keyword lists are expanded via semantic synonym discovery using biomedical ontologies such as PubMed (McEntyre and Lipman, 2001). Next, the **Paragraph Identification Agent** extracts relevant text segments, which are passed to the **Answer Generation Agent**. This module uses few-shot and Chain-of-Thought (CoT) strategies to infer attribute values along with supporting evidence. To ensure factual accuracy, the **Verification Agent** employs a secondary LLM-based evaluator. This agent conducts prompt-based reasoning and graph-based queries ofstructured knowledge databases to assess whether the extracted information is scientifically correct and consistent with external literature. Additionally, each extracted data point is assigned a confidence score (0–100). When the confidence score falls below a predefined threshold (typically 70), the system initiates a regeneration process, filtering out low-quality or weakly relevant text segments to improve the reliability of the output. Then, a **Structuring Agent** formats validated outputs into standardized JSON. This JSON file captures attribute values, supporting evidence, document anchors, and confidence scores. An example of this JSON output is presented in Appendix E. To enable multimodal alignment, the **Cross-modal Alignment Agent** links extracted data from different sources. Matches with similarity scores above 85 are considered reliable.

The final component of the system is a **Human-in-the-loop Validation Module**, serving as a quality control mechanism that enables domain experts

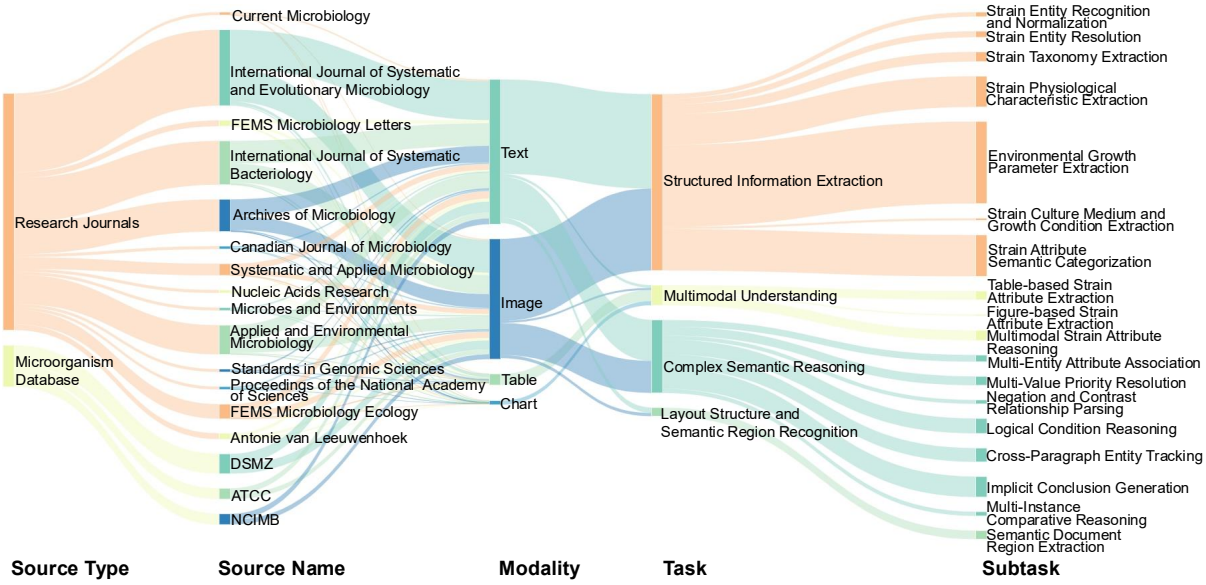


Figure 2: Overview of the QA pair varieties in the MicrobeQuest benchmark based on data source types, specific journal and database sources, supported input modalities, main tasks, and subtasks.

to review and correct extracted information. This platform employs an interactive web interface and implements a rigorous Three-stage validation process:

1. **Preliminary Inspection Stage** The first review group, composed of non-specialist undergraduates, conducts primary validation by checking alignment and consistency between the extracted information and the source text.

2. **Primary Review Stage** The second review group, made up of undergraduates with basic microbiology knowledge, resolves conflicts between initial annotations and AI outputs, with discrepancies highlighted in red to streamline the review process.

3. **Microbiology Expert Final Review Stage** The final review group, composed of microbiology PhDs with practical expertise, conducts thorough quality checks to ensure annotation accuracy and determine database readiness.

Figure 2 illustrates the derivation of 10,176 QA pairs extracted from 127 diverse PDFs using MicrobeCollect, including research journals (e.g., FEMS Microbiology Letters, Applied and Environmental Microbiology) and culture media databases (e.g., DSMZ, ATCC). Additional sources are listed in Appendix A. All the data were divided into 4 main tasks and 18 sub-tasks, with a detailed list provided in Table 2. To ensure data accuracy, the pairs underwent a multi-stage review: an initial inspection by six reviewers, a primary review by two secondary reviewers, and a final review by two microbiology experts.

3.2 Overview of MicrobeQuest Task

As Table 2 shows, MicrobeQuest is structured into four principal capability categories, further subdivided into 18 specialized tasks:

Structured Information Extraction Microbiology research papers contain key strain information, but extracting it is challenging due to its burial in text involving ambiguous terminology, synonyms, and intricate taxonomy. Furthermore, extracting information from such papers often includes condition-dependent features, numerical ranges, and unit conversions. Those challenges demand exact quantitative reasoning and contextual understanding from the IR system. To evaluate this, we propose evaluation tasks focusing on key structured microbial properties: *Strain Entity Recognition and Normalization*, *Strain Entity Resolution*, *Strain Taxonomy Extraction*, *Strain Physiological Characteristic Extraction*, *Environmental Growth Parameter Extraction*, *Strain Attribute Semantic Categorization*, and *Strain Culture Medium and Growth Condition Extraction*.

Multimodal Understanding Unlike general-domain tasks, microbiological data for strain char-

Task	Subtask	Modality	#Task
Structured Information Extraction	Strain Entity Recognition and Normalization	Text/Image	162
	Strain Entity Resolution	Text/Image	221
	Strain Taxonomy Extraction	Text/Image	353
	Strain Physiological Characteristic Extraction	Text/Image	1124
	Environmental Growth Parameter Extraction	Text/Image	3000
	Strain Attribute Semantic Categorization	Text/Image	77
	Strain Culture Medium and Growth Condition Extraction	Text/Image	1516
Multimodal Understanding	Table-based Strain Attribute Extraction	Table	320
	Figure-based Strain Attribute Extraction	Chart	50
	Multimodal Strain Attribute Reasoning	Table/Chart/Text/Image	365
Complex Semantic Reasoning	Multi-Entity Attribute Association	Text/Image	245
	Multi-value Priority Resolution	Text/Image	325
	Negation and Contrast Relationship Parsing	Text/Image	144
	Logical Condition Reasoning	Text/Image	547
	Cross-Paragraph Entity Tracking	Text/Image	505
	Implicit Conclusion Generation	Text/Image	750
	Multi-Instance Comparative Reasoning	Text/Image	156
Layout Structure and Semantic Region Recognition	Semantic Document Region Extraction	Text/Image	316

Table 2: MicrobeQuest encompasses four main tasks and eighteen sub-tasks, totaling 10,176 QA pairs specifically designed for microbiology information retrieval benchmarking.

acterization is highly specialized and interdependent. This data is often presented in multimodal formats like cross-page text, structured tables (e.g., temperature ranges, pH levels), and visual charts (e.g., growth curves, bar charts). Structured tables, for instance, contain condition-sensitive numerical values that require surrounding textual context for proper understanding. Similarly, extracting meaning from charts requires complex reasoning over time and multiple interacting parameters. To assess capability in multimodal understanding, we propose three tasks: *Table-based Strain Attribute Extraction*, *Figure-based Strain Attribute Extraction*, and *Multimodal Strain Attribute Reasoning*.

Complex Semantic Reasoning Academic descriptions of microbial strains involve logical reasoning and implicit knowledge rather than just straightforward facts. Due to dense domain-specific terminology and multi-layered dependencies, the complexity of this area means key information is often implied. Consequently, integrating biochemical, environmental, and strain-specific factors is essential for accurate interpretation. For example, consider the statement: *"When H_2 was replaced by N_2, no growth or methanogenesis occurred unless methanol (50 mM) or acetate (50 mM) was added."* Accurate identification of the growth substrate here involves understanding nested conditionals: the strain does not grow in a nitrogen environment unless specific compounds are added, implicitly suggesting that such supplementation may not be needed under hydrogen. This level of reasoning is common in microbiological literature and exceeds the complexity found in standard IR tasks. To evaluate this capability, we define seven tasks: *Multi-Entity Attribute Association*, *Multi-value Priority Resolution*, *Negation and Contrast Relationship Parsing*, *Logical Condition Reasoning*, *Cross-Paragraph Entity Tracking*, *Implicit Conclusion Generation*, and *Multi-Instance Comparative Reasoning*.

Layout Structure and Semantic Region Recognition Microbiology literature exhibits diverse layouts, ranging from complex multi-column scientific papers with dense tables and figures to simpler single-column records like culture medium sheets. This variability demands models capable of adapting to different document structures, as accurate recognition of key components (such as titles, affiliations, metadata, and references) is crucial for systematic knowledge extraction and serves as a foundation for building structured knowledge graphs. To this end, we introduce the task of *Semantic Document Region Extraction*.

Each MicrobeQuest benchmark task targets distinct challenges for IR within the microbiological literature. Detailed task descriptions are listed in Appendix D.

4 Evaluation

4.1 Benchmark Models

We evaluated the performance of 17 IR models on the MicrobeQuest benchmark with standard prompts (Appendix C) and JSON in-

Task	Subtask	MistralOCR+ Lama-4-Scout-17b-16e-Instruct	PyMuPDF4LLM+ THUDM/GLM-4-32B-0414	MistralOCR+ THUDM/GLM-4-32B-0414	MistralOCR+ Qwen-Coder-Plus	Qwen-Max	GPT-o1
Structured Information Extraction	Strain Entity Recognition and Normalization	0.633	0.610	0.718	0.730	0.678	0.704
	Strain Entity Resolution	0.506	0.562	0.588	0.550	0.511	0.712
	Strain Taxonomy Extraction	0.494	0.458	0.455	0.313	0.342	0.429
	Strain Physiological Characteristic Extraction	0.719	0.757	0.779	0.730	0.752	0.653
	Environmental Growth Parameter Extraction	0.717	0.676	0.665	0.710	0.628	0.699
	Strain Attribute Semantic Categorization	0.650	0.676	0.624	0.701	0.734	0.852
	Strain Culture Medium and Growth Condition Extraction	0.835	0.816	0.797	0.882	0.766	0.759
Multimodal Understanding	Table-based Strain Attribute Extraction	0.605	0.581	0.622	0.615	0.516	0.694
	Figure-based Strain Attribute Extraction	0.414	0.375	0.457	0.426	0.462	0.388
	Multimodal Strain Attribute Reasoning	0.635	0.655	0.656	0.644	0.605	0.689
Complex Semantic Reasoning	Multi-Entity Attribute Association	0.485	0.607	0.568	0.559	0.598	0.685
	Multi-value Priority Resolution	0.676	0.597	0.635	0.703	0.616	0.676
	Negation and Contrast Relationship Parsing	0.844	0.847	0.861	0.893	0.852	0.870
	Logical Condition Reasoning	0.805	0.645	0.800	0.811	0.722	0.675
	Cross-paragraph Entity Tracking	0.637	0.528	0.561	0.515	0.511	0.563
	Implicit Conclusion Generation	0.577	0.485	0.512	0.529	0.592	0.483
	Multi-instance Comparative Reasoning	0.534	0.500	0.552	0.558	0.446	0.455
Layout Structure and Semantic Region Recognition	Semantic Document Region Extraction	0.561	0.906	0.639	0.544	0.940	0.938
Overall F1 Score		0.629	0.627	0.638	0.634	0.626	0.662

Table 3: The F1 Score of top six models on the MicrobeQuest multimodal benchmark. Blue text indicates the open-source models, orange text signifies the closed-source models and red number indicates the best-performances model. The complete list of all model performances is detailed in Appendix F

put(Appendix E). DeepSeek R1 (DeepSeek-AI et al., 2025) results were not included because their API server was consistently busy, preventing the completion of all tasks. Since the benchmark relies on multimodal scientific literature as input, we applied functional enhancements to models lacking native document parsing capabilities. Specifically, professional OCR engines were integrated to convert PDF documents into machine-readable text, ensuring consistent information acquisition conditions across all systems. For language models with native multimodal processing capabilities, we directly invoked their original interfaces to process mixed text-image inputs, thereby preserving the integrity of their architectures. For inference, we adopted a unified strategy that combines COT reasoning and Few-Shot Learning. Designed exemplars and reasoning paths were employed to guide the models through complex knowledge extraction tasks, enabling each system to fully leverage its capacity in handling intricate microbiological information. The list of all models is detailed in Appendix F.

4.2 Implementation Details

We used varied experimental environments to meet the diverse requirements of each model: proprietary commercial models were accessed via their official APIs, while open-source models were retrieved from HuggingFace[1] and deployed according to their parameter scales. Small-scale models were hosted on local servers equipped with NVIDIA RTX 4090 GPUs, whereas computationally intensive large-scale models were executed on high-performance cloud-based GPU clusters to ensure efficient execution of their full computational graphs. The complete list of all model configurations is detailed in Appendix B.

[1] https://huggingface.co/

4.3 Evaluation Metric

To assess the performance of IR models on MicrobeQuest, we employed three widely used metrics: standard F1 score, accuracy, and the Bilingual Evaluation Understudy (BLEU) score (Papineni et al., 2002). In BLEU Score, the Modified Precision$_n$(MP$_n$) measures the accuracy of n-grams in a generated text by comparing them to reference texts, using a minimum count (clipping) to avoid rewarding repetition. This metric is the ratio of the sum of these clipped n-gram counts across all sentences to the total number of n-grams in the generated text. and equation as follows:

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log P_n\right)$$

Where the components are defined as:

- **Brevity Penalty (BP)**: Scales the BLEU score, reducing it for candidate translations that are shorter than the reference translations by comparing the candidate's length (c) to the effective reference length (r).

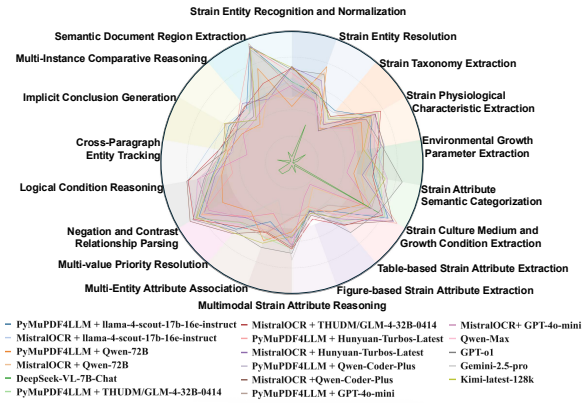


Figure 3: F1 performance comparison of 17 SOTA IR Models across 18 tasks in the MicrobeQuest benchmark.

- **Modified N-gram Precision** (P_n): Quantifies the proportion of n-grams (up to a maximum order N) that are present in both the candidate and reference translations, with counts clipped to prevent over-representation

- w_n: Positive weights for each n-gram precision P_n, typically set uniformly such that $\sum_{n=1}^{N} w_n = 1$ (e.g., $w_n = 1/N$).

4.4 Results

Figure 3 shows the performance overview of all 17 IR methods on MicrobeQuest, and the top six models perform as presented in Table 3. Other model performance results are presented in the Appendix F. Specifically, GPT-o1 proved to be the most effective for structured information extraction and multimodal understanding, while MistralOCR + Qwen-Coder-Plus excelled in complex semantic reasoning. Furthermore, Qwen-Max exhibited the highest proficiency in layout structure and semantic region recognition. Taken together, GPT-o1 achieved the best overall performance among all evaluated models, but the margin over the second-best model was not significant. This outcome reflects the challenges that general-purpose models face due to the lack of domain-specific knowledge, leading to generally lower scores in tasks such as strain resolution and taxonomy extraction. Moreover, interpreting tables and figures via OCR remains difficult, resulting in visual models generally outperforming text-only models on multimodal tasks. For complex reasoning tasks, such as parsing multi-entity and multi-instance information, model performance was also consistently low. Detailed error case analyses are provided in Appendix G.

In general, utilizing OCR preprocessing yielded better outcomes for text-based QA tasks compared to using vision models directly. However, vision models demonstrated distinct advantages in multimodal tasks. Based on these performance profiles and considering cost-effectiveness, we recommend utilizing MistralOCR + THUDM/GLM-4-32B-0414 for microbiology-specialized IR tasks where a readily deployable model is desired.

5 Conclusions

Advanced IR systems facilitate the harnessing of large amounts of high-quality training data, which accelerates AI development in microbiology. However, microbiological data is often embedded within complex, multimodal data (text, images, tables, and charts) scattered across scientific literature and heterogeneous databases. Consequently, progress in the field has been hampered by the lack of specialized benchmarks for evaluating IR system performance and accuracy. Furthermore, the creation of expert-annotated datasets required for such benchmarks is resource-intensive.

To address this critical evaluation gap and facilitate the needs of microbiology-specialized benchmarks, we introduced MicrobeCollect, an expert-in-the-loop platform designed to streamline the acquisition and annotation of multimodal microbiological data. Using this platform, we constructed MicrobeQuest, the first comprehensive, multimodal benchmark specifically tailored for microbiology IR tasks. Developed through collaboration with domain experts, MicrobeQuest encompasses 10,176 question-answer pairs across 18 distinct sub-tasks, targeting essential IR capabilities such as domain-specific knowledge extraction, structured information retrieval, multimodal understanding, and complex semantic reasoning within microbiological literature.

Our extensive evaluation of 17 state-of-the-art IR methods on MicrobeQuest provides crucial baseline performance insights and highlights the varying strengths of different models across microbiology-specific task categories. The results underscore the necessity of domain-specific benchmarks for accurately assessing and advancing IR capabilities in this specialized scientific field.

6 Limitations

This study represents a significant step towards enabling information retrieval in the complex domain of microbiology through the development of the MicrobeCollect platform and the MicrobeQuest benchmark. However, we acknowledge several limitations in the current phase of this study that present opportunities for future work.

Firstly, while the MicrobeCollect tool automates most of the IR process, the necessity for microbiology expert input remains crucial for correcting retrieved information and ensuring the high quality of the resulting annotated data. This dependency introduces a bottleneck, as expert time is valuable and limited, which can constrain the scalability of data preparation. To mitigate this, we believe we can significantly improve our platform by incorporating advanced techniques. For instance, integrating Retrieval-Augmented Generation (RAG) could enhance the initial retrieval accuracy by grounding the model's output in relevant documents, reducing the need for expert correction of factual errors. Similarly, active learning techniques could help prioritize which extracted instances are most uncertain or potentially incorrect, allowing experts to focus their validation efforts more efficiently on the most impactful examples.

Secondly, the size of our current evaluation dataset, comprising 10,176 QA pairs, is constrained by the considerable cost and effort associated with manual annotation and expert validation. While this dataset is carefully curated and domain-specific, a larger dataset would enable more robust evaluation of IR systems and provide a richer resource for training more sophisticated models capable of handling the nuances of microbiology literature. In the future, we plan to address this limitation by exploring the use of Large Language Models (LLMs) to generate additional evaluation QA pairs. Leveraging the generative capabilities of LLMs could allow us to significantly expand the dataset size in a more cost-effective manner. However, this approach will require careful strategies to ensure the accuracy and quality of the synthetically generated data, potentially involving novel LLM-based validation methods or strategic sampling for expert review.

Addressing these limitations in future work will be essential for further advancing AI-driven information retrieval in microbiology and providing more comprehensive resources for the community.

7 Ethics Statement

We all comply with the ACL Ethics Policy[2] in this study. For papers with or without PDFs, we do not provide the original PDFs directly. Instead, we provide a downloader written in Python that users can employ to automatically acquire the raw PDFs if they have the necessary license. Furthermore, we ensured compliance with the licensing agreements associated with these datasets by formally obtaining usage permissions where required. As a result, our research does not raise ethical concerns regarding the use or handling of data.

References

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, and 1 others. 2024. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630(8016):493–500.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. 2018. Ms marco: A human generated machine reading comprehension dataset. *Preprint*, arXiv:1611.09268.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *Preprint*, arXiv:2302.04023.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Hengxing Cai, Xiaochen Cai, Junhan Chang, Sihang Li, Lin Yao, Changxin Wang, Zhifeng Gao, Hongshuai Wang, Yongge Li, Mujie Lin, Shuwen Yang, Jiankun Wang, Mingjun Xu, Jin Huang, Xi Fang, Jiaxi Zhuang, Yuqi Yin, Yaqi Li, Changhong Chen, and 4 others. 2024. Sciassess: Benchmarking llm proficiency in scientific literature analysis. *Preprint*, arXiv:2403.01976.

Jiaxi Cui, Zongjian Li, Yang Yan, Bohua Chen, and Li Yuan. 2023. Chatlaw: Open-source legal large language model with integrated external knowledge bases. *CoRR*.

[2] https://www.aclweb.org/portal/content/acl-code-ethics

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Sandra Dérozier, Robert Bossy, Louise Deléger, Mouhamadou Ba, Estelle Chaix, Olivier Harlé, Valentin Loux, Hélène Falentin, and Claire Nédellec. 2023. Omnicrobe, an open-access database of microbial habitats and phenotypes using a comprehensive text mining and data fusion approach. *PloS one*, 18(1):e0272473.

Carl Edwards, ChengXiang Zhai, and Heng Ji. 2021. Text2mol: Cross-modal molecule retrieval with natural language queries. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 595–607.

Louie Giray. 2023. Prompt engineering with chatgpt: a guide for academic writers. *Annals of biomedical engineering*, 51(12):2629–2633.

Deepak Gupta, Dina Demner-Fushman, William Hersh, Steven Bedrick, and Kirk Roberts. 2024. Overview of trec 2024 biomedical generative retrieval (biogen) track. *Preprint*, arXiv:2411.18069.

Timothy J Hackmann and Bo Zhang. 2021. Using neural networks to mine text and predict metabolic traits for thousands of microbes. *PLoS Computational Biology*, 17(3):e1008757.

A Hoarfrost, A Aptekmann, G Farfañuk, and Y Bromberg. 2022. Deep learning of a bacterial and archaeal universal language of life enables transfer learning and illuminates microbial dark matter. *Nature communications*, 13(1):2606.

Anthony Kay. 2007. Tesseract: an open-source optical character recognition engine. *Linux Journal*, 2007(159):2.

Geewook Kim, Teakgyu Hong, Moonbin Yim, Jeongyeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. 2022. Ocr-free document understanding transformer. *Preprint*, arXiv:2111.15664.

Julia Koblitz, Shuichi Kawashima, Kohei Hidaka, Yoko Okabeppu, Satoshi Onoda, Risa Otsuka, and Natsuko Ichikawa. 2024. Expanding data on cultivation media and microbial traits.

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. Halueval: A large-scale hallucination evaluation benchmark for large language models. *Preprint*, arXiv:2305.11747.

Alejandro Lozano, Jeffrey Nirschl, James Burgess, Sanket Rajan Gupte, Yuhui Zhang, Alyssa Unell, and Serena Yeung-Levy. 2024. μ-bench: A vision-language benchmark for microscopy understanding. *Preprint*, arXiv:2407.01791.

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, Yaofeng Sun, Chengqi Deng, Hanwei Xu, Zhenda Xie, and Chong Ruan. 2024. Deepseek-vl: Towards real-world vision-language understanding. *Preprint*, arXiv:2403.05525.

Wei Lu, Rachel K Luu, and Markus J Buehler. 2025. Fine-tuning large language models for domain adaptation: Exploration of training strategies, scaling, model merging and synergistic capabilities. *npj Computational Materials*, 11(1):84.

Joshua S Madin, Daniel A Nielsen, Maria Brbic, Ross Corkrey, David Danko, Kyle Edwards, Martin KM Engqvist, Noah Fierer, Jemma L Geoghegan, Michael Gillings, and 1 others. 2020. A synthesis of bacterial and archaeal phenotypic trait data. *Scientific data*, 7(1):170.

Johanna McEntyre and David Lipman. 2001. Pubmed: bridging the information gap. *Cmaj*, 164(9):1317–1319.

Eric Nguyen, Michael Poli, Matthew G Durrant, Brian Kang, Dhruva Katrekar, David B Li, Liam J Bartie, Armin W Thomas, Samuel H King, Garyk Brixi, and 1 others. 2024. Sequence modeling and design from molecular to genome scale with evo. *Science*, 386(6723):eado9336.

Matthew A Oberhardt, Raphy Zarecki, Sabine Gronow, Elke Lang, Hans-Peter Klenk, Uri Gophna, and Eytan Ruppin. 2015. Harnessing the landscape of microbial culture media to predict new organism–media pairings. *Nature communications*, 6(1):8493.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Juan Ramos and 1 others. 2003. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, volume 242, pages 29–48. Citeseer.

Derek Reiman, Ahmed Metwally, and Yang Dai. 2017. Using convolutional neural networks to explore the microbiome. In *2017 39th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*, pages 4269–4272. IEEE.

Stephen Robertson, Hugo Zaragoza, and 1 others. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Isabel Schober, Julia Koblitz, Joaquim Sardà Carbasse, Christian Ebeling, Marvin Leon Schmidt, Adam Podstawka, Rohit Gupta, Vinodh Ilangovan, Javad Chamanara, Jörg Overmann, and 1 others. 2025. Bac dive in 2025: the core database for prokaryotic strain data. *Nucleic Acids Research*, 53(D1):D748–D756.

Aditi Singh, Abul Ehtesham, Saket Kumar, and Tala Talaei Khoei. 2025. Agentic retrieval-augmented generation: A survey on agentic rag. *Preprint*, arXiv:2501.09136.

Shamane Siriwardhana, Rivindu Weerasekera, Elliott Wen, Tharindu Kaluarachchi, Rajib Rana, and Suranga Nanayakkara. 2023. Improving the domain adaptation of retrieval augmented generation (rag) models for open domain question answering. *Transactions of the Association for Computational Linguistics*, 11:1–17.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, and 432 others. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Preprint*, arXiv:2206.04615.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models. *Preprint*, arXiv:2104.08663.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine*, 29(8):1930–1940.

Ellen M Voorhees, Donna K Harman, and 1 others. 2005. *TREC: Experiment and evaluation in information retrieval*, volume 63. Citeseer.

David Wadden, Kejian Shi, Jacob Morrison, Aakanksha Naik, Shruti Singh, Nitzan Barzilay, Kyle Lo, Tom Hope, Luca Soldaini, Shannon Zejiang Shen, Doug Downey, Hannaneh Hajishirzi, and Arman Cohan. 2024. Sciriff: A resource to enhance language model instruction-following over scientific literature. *Preprint*, arXiv:2406.07835.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, and 1 others. 2024. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. Bloomberggpt: A large language model for finance. *Preprint*, arXiv:2303.17564.

Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. 2023. The dawn of lmms: Preliminary explorations with gpt-4v(ision). *Preprint*, arXiv:2309.17421.

Hang Yu and Jared R Leadbetter. 2020. Bacterial chemolithoautotrophy via manganese oxidation. *Nature*, 583(7816):453–458.

Haris Zafeiropoulos, Savvas Paragkamian, Stelios Ninidakis, Georgios A Pavlopoulos, Lars Juhl Jensen, and Evangelos Pafilis. 2022. Prego: a literature and data-mining resource to associate microorganisms, biological processes, and environment types. *Microorganisms*, 10(2):293.

Ningyu Zhang, Mosha Chen, Zhen Bi, Xiaozhuan Liang, Lei Li, Xin Shang, Kangping Yin, Chuanqi Tan, Jian Xu, Fei Huang, Luo Si, Yuan Ni, Guotong Xie, Zhifang Sui, Baobao Chang, Hui Zong, Zheng Yuan, Linfeng Li, Jun Yan, and 4 others. 2022. Cblue: A chinese biomedical language understanding evaluation benchmark. *Preprint*, arXiv:2106.08087.

Ruichen Zhang, Shunpu Tang, Yinqiu Liu, Dusit Niyato, Zehui Xiong, Sumei Sun, Shiwen Mao, and Zhu Han. 2025. Toward agentic ai: Generative information retrieval inspired intelligent communications and networking. *Preprint*, arXiv:2502.16866.

Yongshuo Zong, Ismail Elezi, Yongxin Yang, Jiankang Deng, and Timothy Hospedales. 2024. Long-context vision large language models: Empirical insights and a baseline. In *Workshop on Long Context Foundation Models*.

A Data Source

This appendix section presents the sources of literature used in the MicrobeQuest datasets. Figure 4 lists the journals of the papers, and Figure 5 shows the associated microorganism databases.

B Model Configuration

This appendix section presents the models used in our study, as shown in Figure 6. The figure includes each model's name, type (open-source vs. closed-source), description, development environments, and provider.

C Standard Prompt Templates

This appendix section presents the standardized prompt templates used for each task in our benchmark. Each template is designed to elicit specific microbiology-related capabilities from large language models (LLMs).

Name	Description
International Journal of Systematic and Evolutionary Microbiology	The internationally authoritative journal for the classification and evolution of microbial systems is the official platform for the publication of new species, published by the Microbiology Society
International Journal of Systematic Bacteriology	The predecessor of IJSEM, which focused on research in the field of systemic microbiology, was replaced by IJSEM in 2000
Archives of Microbiology	Springer publishes microbiology journals covering multiple fields such as microbial ecology, genetics, biotechnology, and more
Applied and Environmental Microbiology	Published by the American Society for Microbiology (ASM), focusing on environmental and applied microbiology research
Systematic and Applied Microbiology	Published in Microbial Systematics and Applied Research by Elsevier
FEMS Microbiology Letters	The European Microbiology Federation publishes short studies covering a wide range of microbiological fields
Antonie van Leeuwenhoek	Springer Publishing, named after the "Father of Microbiology," publishes research on microbial system classification, ecology, and biodiversity
Current Microbiology	Springer Publishing covers both basic and applied microbiology research
Canadian Journal of Microbiology	Published by Canadian Science Press, dedicated to research in microbial genetics, physiology, ecology, and more
Nucleic Acids Research	Published by Oxford University Press, covering nucleic acid research and bioinformatics resources
Microbes and Environments	Published by the Japanese Society of Microbial Ecology, emphasizing the role of microorganisms in the natural environment
Standards in Genomic Sciences	Focus on genome data publication and standardization processes
Proceedings of the National Academy of Sciences	The Proceedings of the National Academy of Sciences in the United States have strong comprehensiveness and cover multiple disciplines such as microbiology
FEMS Microbiology Ecology	Publish research related to microbial ecology, covering both natural and artificial ecosystems

Figure 4: Overview of Research Journals in MicrobeQuest Benchmark: Names and Details

Name	Description
Deutsche Sammlung von Mikroorganismen und Zellkulturen (DSMZ)	The German Microbial and Cell Bank is one of the world's leading public strain resource centers
American Type Culture Collection (ATCC)	The American Standard Strain Collection is widely used in scientific research and industry
National Collection of Industrial Food and Marine Bacteria (NCIMB)	The National Centre for Industrial, Food and Marine Bacteria in the United Kingdom

Figure 5: Overview of Microorganism Databases in MicrobeQuest Benchmark: Names and Details

C.1 Structured Information Extraction

C.1.1 Strain Entity Recognition and Normalization

This task evaluates a model's ability to identify and standardize microbial strain nomenclature across diverse naming conventions, ensuring consistent strain identification in scientific literature.

Prompt Template

You are a microbiology expert specializing in the accurate identification and normalization of microbial strain names. You will be given a passage of text. Your task is to analyze the content and extract all microbial strain mentions, then normalize them to their standardized forms. Please respond to the following question: `{question}`
Follow the reasoning steps provided below to complete the task: `{note}`
Present your final answer enclosed within the `<Answer>` tags as shown below:
`<Answer>` your_answer_here `</Answer>`

C.1.2 Strain Entity Resolution

This task assesses a model's capacity to determine whether different strain references correspond to the same microbial entity, requiring so-

Name	Type	Description	Env	Company
GPT-4o-mini	closed source	A lightweight vari-ant of GPT-4 optimized for efficient inference and cost-effective deployment. The model we use is GPT-4o-mini	OpenAI API	OpenAI
hunyuan-turbos-latest	closed source	Hunyuan TurboS is a high-speed large language model from Tencent, optimized for low-latency performance in natural language understanding, generation, and code tasks. The model we use is hunyuan-turbos-latest.Using Tencent Hybrid API Interface.	Tencent Hybrid API Interface	Tencent
qwen-coder-plus	closed source	qwen-coder-plus is a code-centric language model from Alibaba's Qwen series, optimized for programming tasks such as code generation, completion, debugging, and multi-language understanding. It is designed to support developers in complex software engineering workflows. The model we use is qwen-coder-plus.	AliCloud QWQ API	Alibaba
Qwen-72B	open source	Qwen-72B is a large-scale language model with 72 billion parameters developed by Alibaba Cloud. It demonstrates strong capabilities in reasoning, multilingual understanding, and instruction following. The model we use is Qwen-72B.	Alibaba Cloud Linux 3.2104 LTS 64-bit, 192 GiB, 300 GiB of space	Alibaba
Qwen-Max	closed source	Qwen-Max is a high-performance model from Alibaba's Qwen series, optimized for complex reasoning, long-context understanding, and multilingual tasks across diverse domains. The model we use is Qwen-Max.	AliCloud QWQ API	Alibaba
GPT-o1	closed source	GPT-o1 is an early iteration of OpenAI's GPT series, focusing on text generation, understanding, and conversational capabilities. It is designed to handle a wide range of tasks, from simple queries to complex reasoning. The model we use is GPT-o1	OpenAI API	OpenAI
Gemini-2.5	closed source	Gemini-2.5 is an advanced version of Google's Gemini model family, excelling in multimodal comprehension. It integrates text, code, image, and audio analysis to deliver sophisticated insights and perform tasks across various domains. The model we use is Gemini-2.5-Pro	Gemini Official API	Google DeepMind
DeepSeek-VL-7B-Chat	open source	DeepSeek-VL is a large language model designed for both visual and textual comprehension, with 7 billion parameters. It excels in multimodal tasks, integrating vision and language understanding to perform complex dialogue and image-related tasks. The model we use is DeepSeek-VL-7B-Chat.	Local configuration deployment of i9 14900KF, 4080s, 32RAM, 2T	Deepseek-AI
Kimi-latest	closed source	Kimi-latest is a high-performance language model optimized for handling large context windows, with a focus on long-form text generation, analysis, and detailed reasoning. It supports tasks such as complex problem-solving and in-depth content creation. The model we use is Kimi-latest-128k	Moonshot official API	Moonshot AI
llama-4-scout-17b-16e-instruct	open source	Llama-4 is an instruction-tuned version of Meta's Llama model, featuring 17 billion parameters. It excels in tasks that require following detailed instructions, including text generation, problem-solving, and contextual reasoning. The model we use is llama-4-scout-17b-16e-instruct.	Alibaba Cloud Linux 3.2104 LTS 64-bit, 192 GiB, 300 GiB of space.	Meta
THUDM/GLM-4-32B-0414	closed source	GLM-4 is a large language model developed by Tsinghua University's THUDM, featuring 32 billion parameters. It is optimized for understanding and generating text in multiple languages, with a focus on high-performance reasoning and complex task execution. The model we use is THUDM/ GLM-4-32B-0414.	silicon-based flow API	ThuDM
MistralOCR	closed source	MistralOCR is an optical character recognition (OCR) model designed to efficiently extract text from scanned documents, images, and PDFs. It combines advanced image processing techniques with natural language understanding to handle complex document structures and deliver accurate text recognition. The model we use is MistralOCR.	Mistral official API	Mistral AI
PyMuPDF4LLM	open source	PyMuPDF4LLM is a specialized language model designed to enhance document processing and analysis. It integrates advanced natural language understanding with PyMuPDF's document handling capabilities, enabling tasks such as content extraction, summarization, and information retrieval from complex documents. The model we use is PyMuPDF4LLM.	Local configuration deployment of i9 14900KF, 4080s, 32RAM, 2T	pymupdf

Figure 6: A Comprehensive Overview of Models: Detailed Insights into Names, Types, Descriptions, Development Environments, and Providers

phisticated understanding of taxonomic relationships and nomenclature evolution.

Prompt Template

You are a microbiology expert specializing in the accurate identification and normalization of microbial strain names. Your expertise includes taxonomic classification, strain naming conventions, and resolving strain synonyms across different nomenclature systems. You will be given a passage of text. Your task is to carefully analyze the content and determine whether different descriptions of a strain refer to the same microbial entity. This includes recognizing when different naming formats, abbreviations, or historical nomenclature may refer to identical strains.
Please respond to the following question: `{question}`

Follow the reasoning steps provided below to complete the task: `{note}`
Present your final answer enclosed within the `<Answer>` tags as shown below:
`<Answer>` your_answer_here `</Answer>`

C.1.3 Strain Taxonomy Extraction

This task aims to extract the taxonomic classification of a microbial strain from scientific text. The goal is to identify the most precise taxonomic information available, including domain, family, genus, etc.

Prompt Template

You are a microbiology expert with specialized knowledge in the accurate identification and reasoning of microbial taxonomic information. Your expertise includes taxo-

nomic classification, phylogenetic relationships, and nomenclatural interpretation.
I will provide you with a passage of text. Your task is to carefully analyze the content and determine the taxonomic classification of the strain, including genus, species, and other relevant ranks if available.
Please answer the following question: `{question}`
Follow the reasoning steps provided below to complete the task: `{note}`
Present your final answer within the `<Answer>` tags as shown below:
`<Answer>` your_answer_here `</Answer>`

C.1.4 Strain Physiological Characteristic Extraction

This task evaluates a model's ability to precisely extract key physiological traits of microbial strains from scientific texts, including morphological features and biochemical properties essential for strain identification.

Prompt Template

You are a microbiology expert with specialized knowledge in identifying and extracting microbial physiological traits. Your expertise includes bacterial morphology, biochemical test interpretation, and cellular characteristic analysis across diverse microbial species. You will be given a passage of text. Your task is to analyze the content and extract key physiological characteristics of strains, such as Gram-staining results, motility, oxygen requirements, cell morphology, and strain types.
Please respond to the following question: `{question}`
Follow the reasoning steps provided below to complete the task: `{note}`
Present your final answer enclosed within the `<Answer>` tags as shown below:
`<Answer>` your_answer_here `</Answer>`

C.1.5 Environmental Growth Parameter Extraction

This task focuses on evaluating the model's ability to extract specific environmental factors required for microbial growth (e.g., temperature, pH) from unstructured scientific literature.

Prompt Template

You are an expert in microbiology with specialized knowledge in the precise determination of microbial growth requirements and environmental tolerances. Your expertise includes understanding how environmental factors affect microbial metabolism, reproduction, and survival. You will be given a passage of text. Your task is to carefully analyze the content and extract key environmental parameters required for optimal strain growth, including temperature ranges, pH tolerance, salinity requirements, and oxygen availability preferences. Please answer the following question: `{question}` Follow the reasoning steps provided below to complete the task: `{note}` Present your final answer within the `<Answer>` tags as shown below:
`<Answer>` your_answer_here `</Answer>`

C.1.6 Strain Attribute Semantic Categorization

This task evaluates a model's ability to classify extracted microbial strain attributes into standardized semantic categories, requiring a deep understanding of microbiological terminology.

Prompt Template

You are an expert in microbiology with specialized knowledge in the precise identification and categorization of strain attributes extracted from text into standardized semantic categories, such as growth environment types.
I will provide you with a passage of text. Your task is to carefully analyze the content and categorize identified strain attributes according to established taxonomic and semantic frameworks.
Please answer the following question: `{question}`
Follow the reasoning steps provided below to complete the task: `{note}`

Present your final answer within the <Answer> tags as shown below: <Answer> your_answer_here </Answer>

C.1.7 Strain Culture Medium and Growth Condition Extraction

This task assesses a model's ability to identify and structure detailed information about culture media components and specific cultivation conditions required for successful microbial growth from technical descriptions.

Prompt Template

You are an expert in microbiology with specialized knowledge in the formulation of culture media and the optimization of growth conditions for diverse microbial strains. Your expertise includes media composition, environmental requirements for cultivation, and techniques to promote the growth of fastidious organisms.
You will be given a passage of text. Your task is to analyze the content and extract detailed descriptions of the culture medium components and growth conditions required for the strain, including key ingredients, specific environmental parameters, and procedural steps.
Please answer the following question: {question}
Follow the reasoning steps provided below: {note}
Provide your final answer within the <Answer> tags as shown: <Answer> your_answer_here </Answer>

C.2 Multimodal Understanding

C.2.1 Table-based Strain Attribute Extraction

This task evaluates a model's capability to accurately interpret and extract structured information from complex tabular data in microbiological literature, requiring sophisticated pattern recognition and relationship inference.

Prompt Template

You are an expert in microbiology with specialized knowledge in the precise interpretation and analysis of tabular microbiological data. Your expertise includes understanding complex taxonomic tables, growth characteristic matrices, and comparative strain property charts used in research publications and laboratory reports.
I will provide you with a passage of text. Your task is to carefully analyze the content and identify and extract critical information from tabular data, addressing challenges such as header interpretation, cell content association, data relationship inference, and handling of missing or partial data in microbial strain documentation.
Please answer the following question: {question}
Follow the reasoning steps provided below to complete the task: {note}
Present your final answer within the <Answer> tags as shown below: <Answer> your_answer_here </Answer>

C.2.2 Figure-based Strain Attribute Extraction

This task tests a model's ability to derive meaningful information from descriptions of graphical representations in microbiology research, including growth curves, metabolic pathways, and microscopy image analyses.

Prompt Template

You are an expert in microbiology with specialized knowledge in the precise interpretation of graphical data representing microbial properties and behaviors. Your expertise includes analyzing growth curves, metabolic pathway diagrams, microscopy image interpretations, and phylogenetic trees commonly found in microbiological research.
I will provide you with a passage of text. Your task is to carefully analyze the content and identify and extract critical information from graphical representations, including curve trend analysis, data point comparison,

legend and axis interpretation, and quantitative result extraction from visual data related to microbial strains.
Please answer the following question: {question}
Follow the reasoning steps provided below to complete the task: {note}
Present your final answer within the <Answer> tags as shown below: <Answer> your_answer_here </Answer>

C.2.3 Multimodal Strain Attribute Reasoning

This task evaluates a model's capacity to integrate and synthesize information about microbial strains across multiple presentation formats, requiring sophisticated cross-modal verification and complementary information processing.

Prompt Template

You are an expert in microbiology with specialized knowledge in the precise integration and synthesis of multimodal microbiological information. Your expertise includes correlating textual descriptions with tabular data and graphical evidence to form comprehensive understandings of microbial characteristics, behaviors, and classifications.
I will provide you with a passage of text. Your task is to carefully analyze the content and integrate and reason across information from text, tables, and images, performing cross-modal verification, complementary information synthesis, and complex inferencing based on multi-source data about microbial strains.
Please answer the following question: {question}
Follow the reasoning steps provided below to complete the task: {note}
Present your final answer within the <Answer> tags as shown below: <Answer> your_answer_here </Answer>

C.3 Complex Semantic Reasoning

C.3.1 Multi-Entity Attribute Association

This task assesses a model's ability to correctly assign attributes to their corresponding microbial entities within complex texts, requiring advanced coreference resolution and entity relationship tracking.

Prompt Template

You are an expert in microbiology with specialized knowledge in the precise identification and association of attributes with their corresponding microbial entities. Your expertise includes resolving complex referential relationships in scientific literature, disambiguating between similar strains, and tracking attribute assignments across dense technical descriptions.
I will provide you with a passage of text. Your task is to carefully analyze the content and identify attributes corresponding to multiple entities within long paragraphs and accurately align attributes to their correct subjects, especially when subjects are omitted or ambiguously referenced in microbiological contexts.
Please answer the following question: {question}
Follow the reasoning steps provided below to complete the task: {note}
Present your final answer within the <Answer> tags as shown below: <Answer> your_answer_here </Answer>

C.3.2 Multi-value Priority Resolution

This task evaluates a model's capability to select the most appropriate value when confronted with multiple conflicting measurements for the same microbial property, requiring contextual reasoning about experimental reliability.

Prompt Template

You are an expert in microbiology with specialized knowledge in evaluating and prioritizing conflicting or multiple reported values for microbial properties. Your expertise includes understanding experimental context, methodological reliability, and standardized reporting systems for strain characteristics.
I will provide you with a passage of text. Your task is to carefully analyze the content and select the most contextually appropri-

ate and semantically prioritized value when multiple candidate values are associated with the same property, considering factors such as experimental conditions, measurement methods, and scientific consensus.
Please answer the following question: {question}
Follow the reasoning steps provided below to complete the task: {note}
Present your final answer within the <Answer> tags as shown below: <Answer> your_answer_here </Answer>

C.3.3 Negation and Contrast Relationship Parsing

This task measures a model's ability to accurately interpret complex linguistic structures involving negation and contrastive relationships in microbiological contexts, essential for extracting factually correct information.

Prompt Template

You are an expert in microbiology with specialized knowledge in the precise interpretation of complex linguistic structures describing microbial properties. Your expertise includes analyzing negation patterns, contrastive relationships, and exception clauses in scientific literature to extract accurate factual information.
I will provide you with a passage of text. Your task is to carefully analyze the content and parse and interpret negation and contrastive relationships to accurately capture the intended factual meaning from complex expressions about microbial strains, their properties, and behaviors.
Please answer the following question: {question}
Follow the reasoning steps provided below to complete the task: {note}
Present your final answer within the <Answer> tags as shown below: <Answer> your_answer_here </Answer>

C.3.4 Logical Condition Reasoning

This task evaluates a model's capacity to apply logical frameworks and conditional reasoning to microbiological information, requiring the ability to process complex if-then relationships and dependency chains.

Prompt Template

You are an expert in microbiology with specialized knowledge in the application of logical frameworks to microbiological data interpretation. Your expertise includes understanding complex conditional relationships in experimental designs, metabolic pathways, and growth requirements.
I will provide you with a passage of text. Your task is to carefully analyze the content and infer conclusions based on multiple conditional statements and constraints, including conditional dependencies, "if-then" structures, and contextual logic chains relevant to microbial characteristics and behaviors.
Please answer the following question: {question}
Follow the reasoning steps provided below to complete the task: {note}
Present your final answer within the <Answer> tags as shown below: <Answer> your_answer_here </Answer>

C.3.5 Cross-Paragraph Entity Tracking

This task assesses a model's ability to maintain coherent tracking of microbial entities across multiple paragraphs, requiring sophisticated coreference resolution and information integration across distributed contexts.

Prompt Template

You are an expert in microbiology with specialized knowledge in the coherent integration of distributed information about microbial entities. Your expertise includes maintaining entity consistency across complex research papers, tracking strain references across multiple experimental sections, and resolving co-reference in technical writing.
I will provide you with a passage of text. Your task is to carefully analyze the content and track entities across multiple paragraphs to integrate fragmented information and ensure consistent entity-level under-

standing of microbial strains and their properties.
Please answer the following question: {question}
Follow the reasoning steps provided below to complete the task: {note}
Present your final answer within the <Answer> tags as shown below: <Answer> your_answer_here </Answer>

C.3.6 Implicit Conclusion Generation

This task evaluates a model's ability to derive scientifically sound inferences about microbial properties that are not explicitly stated but logically follow from the provided information.

Prompt Template

You are an expert in microbiology with specialized knowledge in inferential reasoning based on incomplete microbiological data. Your expertise includes drawing scientifically sound conclusions from partial evidence, understanding implied relationships in research findings, and extrapolating valid inferences from experimental results.
I will provide you with a passage of text. Your task is to carefully analyze the content and infer logically valid conclusions that are not explicitly stated in the text, by synthesizing contextual clues, conditions, and implied relationships about microbial strains and their properties.
Please answer the following question: {question}
Follow the reasoning steps provided below to complete the task: {note}
Present your final answer within the <Answer> tags as shown below: <Answer> your_answer_here </Answer>

C.3.7 Multi-Instance Comparative Reasoning

This task measures a model's capability to perform comparative analyses across different experimental instances in microbiology, requiring the ability to identify patterns, differences, and meaningful relationships across complex contexts.

Prompt Template

You are an expert in microbiology with specialized knowledge in the comparative analysis of microbial entities across different experimental conditions. Your expertise includes identifying meaningful patterns across multiple experiments, understanding significance in comparative studies, and drawing conclusions from parallel or contrasting results.
I will provide you with a passage of text. Your task is to carefully analyze the content and perform comparative analysis across different experimental instances or groups to derive conclusions based on observed differences, similarities, or relative outcomes in microbial behavior, growth patterns, or metabolic activities.
Please answer the following question: {question}
Follow the reasoning steps provided below to complete the task: {note}
Present your final answer within the <Answer> tags as shown below: <Answer> your_answer_here </Answer>

C.4 Layout Structure and Semantic Region Recognition

C.4.1 Semantic Document Region Extraction

This task evaluates a model's ability to identify and extract structured information from specific regions within scientific literature, requiring understanding of document architecture and semantic organization in microbiological publications.

Prompt Template

You are an expert in microbiology with specialized knowledge in the precise identification and extraction of structured information from scientific literature. Your expertise includes recognizing standardized document components, understanding scientific publication formats, and extracting semantically meaningful sections from research papers.
I will provide you with a passage of text. Your task is to carefully analyze the content and accurately identify and extract

key semantic regions within scientific literature—such as titles, author names, institutional affiliations, abstracts, introduction sections, methodology descriptions, results, discussions, figure captions, table headers, acknowledgments, and references—maintaining their hierarchical relationships and contextual significance.
Please answer the following question: `{question}`
Follow the reasoning steps provided below to complete the task: `{note}`
Present your final answer within the `<Answer>` tags as shown below: `<Answer>` your_answer_here `</Answer>`

D Subtask Description

In this appendix, we provide detailed definitions and representative examples for all evaluation tasks introduced in the main paper, including sample inputs, expected outputs, and evaluation criteria.

D.1 Structured Information Extraction

D.1.1 Strain Entity Recognition and Normalization

This task focuses on identifying and normalizing standard strain entity names from a given passage. The goal is to extract the correct strain name in its canonical form based on context.

- **Task Input:** A passage of text containing information about a microbial strain entity
- **Task Output:** The normalized strain entity name that accurately answers the question
- **Evaluation Metric:** Exact-match F1 score between the predicted and reference answers

Input:

Based on the results from phylogenetic, morphological, and protein analyses, we conclude that the novel strain represents a novel species of the genus *Methanocaldococcus*, for which the name *Methanocaldococcus villosus* sp. nov. is proposed (type strain KIN24-T80^T = DSM 22612^T = JCM 16315^T).

Question:

`"What is the full name of a strain (species+strain)?"`

Expected Answer:

`<Answer> Methanocaldococcus villosus KIN24-T80 </Answer>`

D.1.2 Strain Entity Resolution

This task focuses on determining whether different mentions of strain identifiers within a given context refer to the same microbial entity.

- **Task Input:** A passage of text containing information about a microbial strain entity
- **Task Output:** A binary `True` or `False` answer indicating whether the given strain mentions refer to the same entity
- **Evaluation Metric:** Exact-match accuracy score between predicted and reference answers

Input:

Based on the results from phylogenetic, morphological, and protein analyses, we conclude that the novel strain represents a novel species of the genus *Methanocaldococcus*, for which the name *Methanocaldococcus villosus* sp. nov. is proposed (type strain KIN24-T80^T = DSM 22612^T = JCM 16315^T).

Question:

`"Is strain KIN24-T80 the same strain entity as strain JCM 16315?"`

Expected Answer:

`<Answer> True </Answer>`

D.1.3 Strain Taxonomy Extraction

This task focuses on extracting the full taxonomic lineage of a given microbial strain, including domain, phylum, class, order, family, genus, and species, as available in the input context.

- **Task Input:** A textual passage that contains taxonomic information about a microbial strain, along with by a specific question
- **Task Output:** The explicit taxonomic classification of the strain corresponding to the question scope (e.g., genus, species)

- **Evaluation Metric:** Exact match accuracy between the predicted and reference answers

Input:

A novel chemolithoautotrophic, hyperthermophilic methanogen was isolated from a submarine hydrothermal system at the Kolbeinsey Ridge, north of Iceland. Based on its 16S rRNA gene sequence, the strain belongs to the order *Methanococcales* within the genus *Methanocaldococcus*, with approximately 95% sequence similarity to *Methanocaldococcus jannaschii* as its closest relative.

Question:

`"What is the genus of the microorganism for the strain KIN24-T80?"`

Expected Answer:

`<Answer> Methanocaldococcus </Answer>`

D.1.4 Strain Physiological Characteristic Extraction

This task focuses on extracting physiological characteristics of microbial strains from a given passage of text. The focus is on intrinsic properties of the strain, such as Gram reaction, motility, and cell morphology.

- **Task Input:** A passage of text describing physiological traits of a microbial strain, along with a specific question
- **Task Output:** The description or value of the strain's physiological trait based on the question scope
- **Evaluation Metric:** Exact-match F1 score between the predicted and reference answers

Input:

Cells of the novel organism stained Gram-negative and appeared as regular to irregular cocci possessing more than 50 polar flagella. These cell appendages mediated not only motility but also adherence to abiotic surfaces and the formation of cell–cell contacts.

Question:

`"What is the Gram reaction for the strain KIN24-T80?"`

Expected Answer:

`<Answer> Gram-negative </Answer>`

D.1.5 Environmental Growth Parameter Extraction

This task focuses on extracting environmental growth parameters of microbial strains from a given passage of text. The emphasis is on growth conditions such as temperature, pH, salinity (NaCl concentration), and required chemical components or elements for growth.

- **Task Input:** A passage of text describing the environmental growth conditions of a microbial strain, along with a specific question
- **Task Output:** The strain's growth parameter or required condition based on the question
- **Evaluation Metric:** Exact-match F1 score between the predicted and reference answers

Input:

Hence, all further experiments were performed at the optimal growth temperature of 80 C. To ascertain the pH dependence of the organism, the pH of the medium was adjusted with diluted sulphuric acid or sodium hydroxide, as indicated above, without the usage of additional buffers. Growth was observed between pH 5.5 and 7.0, with an optimum at pH 6.5; no growth was detected at or below pH 5.0 or at and above pH 7.5. Different amounts of NaCl were added to the culture medium (MGG medium prepared without NaCl) to determine the optimum growth rate with regard to salt concentration. The minimal requirement for growth was 0.5% (w/v) NaCl, the upper limit was 5.5% (w/v) NaCl, and the optimum was 2.5% (w/v) NaCl. The minimum doubling time for growth of strain KIN24-T80^{T} under optimal conditions was 45 min.

Question:

`"What is the minimum NaCl concentration required for growth of the strain KIN24-T80?"`

Expected Answer:

`<Answer> 0.5% </Answer>`

D.1.6 Strain Attribute Semantic Categorization

This task focuses on identifying and categorizing strain attributes related to their growth environments into standardized semantic categories (e.g., geothermal, marine, terrestrial, halophilic, thermophilic).

- **Task Input:** A passage of text describing the growth environment or habitat of a microbial strain, along with a specific question
- **Task Output:** The categorized description of the strain's typical growth environment
- **Evaluation Metric:** Exact-match F1 score between the predicted and reference answers

Input:

The results of phenotypic characterization confirm the affiliation of KIN24-T80^T to the genus *Methanocaldococcus*. Nevertheless, 16S rRNA gene sequence analysis in combination with the unique whole-cell protein SDS-PAGE pattern proved its distinctiveness from any previously described species. Based on the data presented herein, strain KIN24-T80^T represents a novel species, for which the name *Methanocaldococcus villosus* sp. nov. is proposed. Herewith, we describe the first hyperthermophilic *Methanocaldococcus* species isolated from a shallow submarine hydrothermal system.

Question:

`"In what specific habitat or environment does this organism typically grow for the strain KIN24-T80?"`

Expected Answer:

`<Answer>Geothermal</Answer>`

D.1.7 Strain Culture Medium and Growth Condition Extraction

This task focuses on identifying and structuring descriptions of culture medium components and cultivation conditions required for strain growth.

- **Task Input:** A passage of text describing the culture medium and associated question
- **Task Output:** A structured answer to the question about the culture medium composition or growth condition
- **Evaluation Metric:** Exact-match F1 score between the predicted and reference answers

Input:

"Microorganisms DSMZ. PFENNIG'S MEDIUM I consists of multiple solutions. Solution A contains calcium chloride dihydrate (0.25 g), yeast extract (0.25 g), and distilled water (460.00 ml). For marine or estuarine isolates, add 100.0 g NaCl and increase magnesium sulfate heptahydrate to 15.0 g. Solution B includes sodium sulfide nonahydrate (2.00 g) in 135.00 ml distilled water. Solution C is prepared with sodium bicarbonate (1.50 g) in 50.00 ml water, bubbled with CO_2 and filter sterilized. Solution D contains resazurin (0.1%, 0.5 ml) in 450.00 ml distilled water. Solution E includes ammonium chloride (0.35 g), ammonium acetate (0.25 g), pyruvic acid sodium salt (0.25 g), dextrose (0.25 g), magnesium sulfate heptahydrate (0.50 g), potassium chloride (0.35 g), potassium phosphate monobasic (0.35 g), trace element solution SL-12 B (1.00 ml), and distilled water (25 ml), and is filter sterilized. Solution F consists of vitamin B_{12} (0.01 g) in 100.00 ml distilled water, filter sterilized. The trace element solution SL-12 B includes distilled water (1000.00 ml), Na_2-EDTA (3.00 g), $FeSO_4{\cdot}7H_2O$ (1.10 g), $CoCl_2{\cdot}6H_2O$ (190.00 mg), $MnCl_2{\cdot}2H_2O$ (50.00 mg), $ZnCl_2$ (42.00 mg), $NiCl_2{\cdot}6H_2O$ (24.00 mg), $Na_2MoO_4{\cdot}2H_2O$ (18.00 mg), H_3BO_3 (300.00 mg), and $CuCl_2{\cdot}2H_2O$ (2.00 mg), adjusted to pH 6.0. Solutions D, C, and E are mixed, bubbled with CO_2 in an ice bath under sterile conditions, and 50 ml is added to each bottle of solution A. Before use, add 4 ml solution B and 0.1 ml solution F. The final pH is adjusted to 7.1–7.3 using filter-sterilized 1 M Na_2CO_3. The medium is distributed into sterile, nitrogen-gassed screw-cap tubes. During the first 24 hours, iron precipitates as black flocks; no

other sediment should appear. Periodic supplementation with neutralized 3% sodium sulfide solution is required. The sulfide solution is made with $Na_2S{\cdot}9H_2O$ (3.00 g) in 100.00 ml distilled water, bubbled with nitrogen, autoclaved, and pH-adjusted to ~7.0 with sterile 2 M H_2SO_4. A yellow color indicates a drop to pH ~8. The solution is stirred continuously to avoid elemental sulfur precipitation, and the final solution should be clear and yellow."

Question:

"What components are required in the culture medium?"

Expected Answer:

<Answer>Calcium chloride, yeast extract, sodium sulfide, sodium bicarbonate, resazurin, ammonium chloride, ammonium acetate, pyruvate, dextrose, magnesium sulfate, potassium chloride, potassium phosphate, vitamin B12, trace elements (including EDTA, $FeSO_4$, $CoCl_2$, $MnCl_2$, $ZnCl_2$, $NiCl_2$, MoO_4, BO_3, $CuCl_2$)</Answer>

D.2 Multimodal Understanding

D.2.1 Table-based Strain Attribute Extraction

This task aims to extract specific attribute information of microbial strains from tabular data. It addresses challenges such as header interpretation, cell value association, multi-attribute relationship inference, and handling of missing or incomplete information.

- **Task Input:** A table describing multiple strain characteristics (e.g., physiological traits, growth conditions)
- **Task Output:** Structured extraction of target strain attributes in response to a specific question
- **Evaluation Metric:** Exact-match F1 score between the predicted and reference answers

Input:

Table 1. Comparative characteristics of strain B10^T, '*Candidatus* Aciduliprofundum boonei' and *Methanosphaera stadtmanae*

Taxa: 1, strain B10^T (data from this study); 2, '*Candidatus* Aciduliprofundum boonei' DSM 19572 (Reysenbach *et al.*, 2006); 3, *Methanosphaera stadtmanae* DSM 3091^T (Miller & Wolin, 1985). ND, No data available.

Characteristic	1	2	3
Source	Human faeces	Hydrothermal vent field	Human faeces
Morphology	Cocci, as single cells	Pleiomorphic cocci	Cocci, as pairs or tetrads
Diameter (μm)	0.7–1.0	0.6–1.0	1.0–1.2
Motility	Non-motile	Motile	Non-motile
Optima for growth:			
Temperature (°C)	37	70	37
pH	7.6	4.5	6.5–6.9
NaCl (%, w/v)	1.0	ND	ND
DNA G+C content (mol%)	59.93	39	25.8
Cell-wall ultrastructure	Thin electron-dense layer and thick transparent layer	Thick electron-dense layer	Thick electron-dense layer
Substrates for methane production	H_2 + methanol	None	H_2 + methanol

Question:

"What is the optimal NaCl concentration for the strain B-10^T?"

Expected Answer:

<Answer>1%</Answer>

D.2.2 Figure-based Strain Attribute Extraction

This task aims to extract specific attribute information of microbial strains from figure-based data. It addresses challenges such as visual interpretation, pattern recognition, multi-attribute relationship inference, and handling of incomplete or ambiguous visual information.

- **Task Input:** A figure illustrating various strain characteristics (e.g., physiological traits, growth conditions)
- **Task Output:** Structured extraction of target strain attributes in response to a specific question
- **Evaluation Metric:** Exact-match F1 score between the predicted and reference answers

Input:

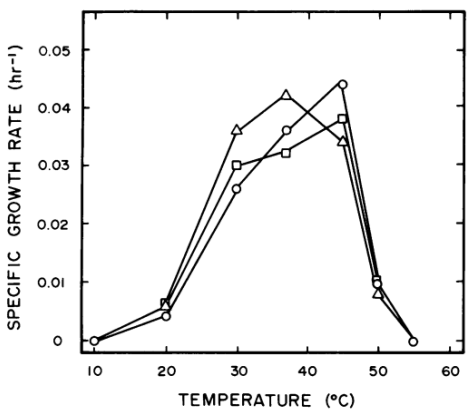


FIG. 2. Relationship between growth rate and incubation temperature for oil field *Methanobacterium* sp. Symbols: ○, strain ivanov; △, strain kuznetsov; □, strain omeliansky.

Question:

"What is the optimal growth temperature for the strain ivanov?"

Expected Answer:

<Answer>46</Answer>

D.2.3 Multimodal Strain Attribute Reasoning

This task focuses on integrating and reasoning across information from text, tables, and images. The information may be present in both the text and the figures, and the model needs to effectively combine multimodal data to answer questions.

- **Task Input:** A passage of text and a table/figure describing the culture medium and associated question.
- **Task Output:** A structured answer to the question about the culture medium composition or growth condition.
- **Evaluation Metric:** Exact-match F1 score between the predicted and reference answers.

Input:

Text: "An autotrophic thermophilic motile coccoid methanogen was isolated from geothermally heated sea sediments near Naples, Italy. Growth occurs on H_2/CO_2 and on formate between 30 and 70°C, with an optimum at 65°C. The optimal doubling time is only 55 minutes. The NaCl concentration ranges from 1.3% to 8.3%, with an optimum around 4%. By its G + C content of 31.3 mol%, its subunit envelope, and by DNA-RNA hybridization, the new isolate is clearly defined as a member of the genus Methanococcus. We name it *Methanococcus thermolithotrophicus SN-1*."

Figure:

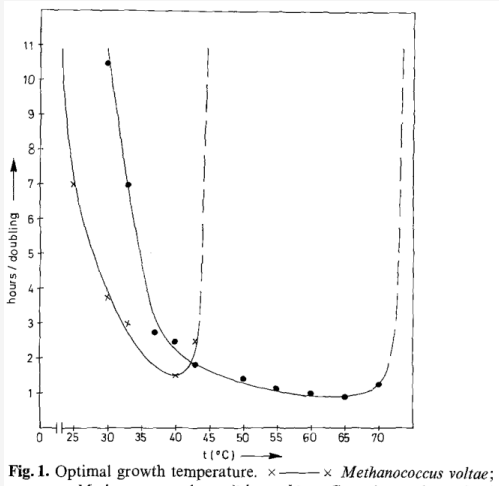


Fig. 1. Optimal growth temperature. ×——× *Methanococcus voltae*; ●——● *Methanococcus thermolithotrophicus*. Growth was determined several times during the exponential phase by $O.D._{578}$-measurement. The hours/doubling were calculated from the slopes of the growth curves (not shown)

Question:

"What is the lower bound of the optimal growth temperature range for *Methanococcus thermolithotrophicus SN-1*?"

Expected Answer:

<Answer>65</Answer>

D.3 Complex Semantic Reasoning

D.3.1 Multi-Entity Attribute Association

This task aims to extract the correct attribute value for a specific microbial entity mentioned in the question from a passage that includes multiple entities and their associated attributes. The model must accurately resolve entity-level ambiguity and ensure that the extracted answer corresponds precisely to the target entity, not to other co-mentioned entities.

- **Task Input:** A passage containing multiple microbial strains, each associated with different attributes, and a question targeting a specific entity's attribute.
- **Task Output:** The exact attribute value (e.g., substrate, temperature, compound name) associated with the target entity mentioned in the question.
- **Evaluation Metric:** Exact-match F1 score between the predicted and reference answers.

Input:

Our results seem to indicate that, in rice fields, *Methanobacterium* spp. are

mostly responsible for CH_4 production from H_2/CO_2, and *Methanosarcina* spp. for CH_4 production from acetate. *Jannaschii* its closest relative.

Question:

"Is acetate supporting the growth for the strain *Methanosarcina* spp.?"

Expected Answer:

<Answer> true </Answer>

D.3.2 Multi-value Priority Resolution

This task focuses on selecting the most appropriate attribute value when multiple candidate values are mentioned throughout the document. The model needs to resolve conflicts or prioritize among values based on contextual cues or scientific conventions.

- **Task Input:** A passage containing multiple candidate values for a given attribute
- **Task Output:** The most appropriate and contextually valid attribute value
- **Evaluation Metric:** Exact-match F1 score between the predicted and reference answers.

Input:

These isolates (strains FDF-17 [T = type strain], FDF-2, SF-2, Ret-1, SD-1, and Cas-1) grew on media containing methanol and mono-, di-, and trimethylamines as catabolic substrates, but not on media containing dimethyl sulfide, methane thiol, H_2, formate, or acetate. Other cultures containing methanol as the catabolic substrate and inoculated in the same way also formed methane, but cultures in media containing acetate, H_2, formate, propionate, butyrate, lactate, or cellulose did not form methane.

Question:

"What substrates do not support the organism's growth for the strain FDF-17?"

Expected Answer:

<Answer>acetate, H_2, formate, propionate, butyrate, lactate, cellulose</Answer>

D.3.3 Negation and Contrast Relationship Parsing

This task focuses on identifying and interpreting negation and contrast relationships within scientific descriptions. The goal is to accurately determine whether specific conditions, substances, or attributes are positively or negatively associated with the target strain.

- **Task Input:** A scientific passage containing multiple statements about strain behavior under different conditions
- **Task Output:** The exact attribute value or Boolean answer that correctly reflects the negated or contrasted relationship
- **Evaluation Metric:** Exact-match F1 score between the predicted and reference answers.

Input:

To test the ability of the isolate to utilize energy sources other than H_2, the medium was prepared with a gas phase of N_2/CO_2 (250 kPa, 80:20, v/v), and the following substrates were added separately to final concentrations of 0.1% (w/v): acetate, formate, methanol, pyruvate, and yeast extract. No growth could be detected over a period of 3 days by phase-contrast microscopy. Therefore, the strain was considered to grow exclusively by reduction of CO_2 using H_2 as an electron donor, like all members of the genus *Methanocaldococcus* with validly published names (Jones et al., 1983; Jeanthon et al., 1998, 1999; L'Haridon et al., 2003).

Question:

"Is N_2 supporting the growth for the strain *Methanosarcina* spp.?"

Expected Answer:

<Answer>false</Answer>

D.3.4 Logical Condition Reasoning

This task focuses on identifying and interpreting logical conclusions derived from multiple conditional statements and constraints. These may

include "if-then" structures, condition-dependent outcomes, and contextual logic chains commonly found in scientific reasoning.

- **Task Input:** A scientific passage containing conditionally structured information or logical dependencies.
- **Task Output:** The exact attribute value or Boolean answer that accurately reflects the inference drawn from the stated conditions.
- **Evaluation Metric:** Exact-match F1 score between the predicted and reference answers.

Input:

To test whether the organism could grow without hydrogen, cultures were prepared with a gas phase of N_2/CO_2 (80:20) and supplemented separately with acetate, formate, methanol, pyruvate, and yeast extract. No growth was observed over 3 days. Therefore, the strain is considered to grow exclusively via CO_2 reduction using H_2 as the electron donor.

Question:

"What substrates support the organism's growth?"

Expected Answer:

<Answer>H_2/CO_2 </Answer>

D.3.5 Cross-Paragraph Entity Tracking

This task focuses on identifying and interpreting entity attributes that span across multiple paragraphs, requiring the model to track and connect relevant entities and their associated properties across different sections of the text.

- **Task Input:** A passage containing information spread across multiple paragraphs, where entities and their attributes need to be identified and linked correctly.
- **Task Output:** The exact attribute value or Boolean answer that accurately reflects the inference drawn from the connections between the entities and their attributes across the paragraphs.
- **Evaluation Metric:** Exact-match F1 score between the predicted and reference answers.

Input:

In order to investigate the growth conditions for *Methanocaldococcus* species, MGG medium was prepared with various substrates, including acetate, H_2, and formate. The organism showed significant growth when acetate and H_2 were present, but no growth was observed with formate as the sole substrate. The strains were incubated at a constant temperature of 37°C for a period of 7 days.

Question:

"What substrates support the organism's growth?"

Expected Answer:

<Answer>acetate,H_2</Answer>

D.3.6 Implicit Conclusion Generation

This task focuses on identifying and interpreting implicit conclusions based on the provided scientific context, where conclusions are inferred from the given information rather than explicitly stated.

- **Task Input:** A passage containing information that indirectly leads to a conclusion.
- **Task Output:** The exact attribute value or Boolean answer that accurately reflects the implicit inference drawn from the passage.
- **Evaluation Metric:** Exact-match F1 score between the predicted and reference answers.

Input:

Selective enrichment culture techniques were employed to obtain mixed cultures of methanogenic rods and sarcina from surface flooding waters and deep subsurface (-1650 m) oil-bearing sedimentary rocks and formation waters sampled from an old oil field in the U.S.S.R. previously reported to display active biological methanogenesis. The methanogens were selectively isolated as colonies on agar petri dishes that were incubated in a novel container.

Question:

"What is the source for the strain?"

Expected Answer:

<Answer>oil-bearing

```
sedimentary rocks, formation
waters</Answer>
```

D.3.7 Multi-Instance Comparative Reasoning

This task focuses on conducting a comparative analysis across different experimental conditions or groups to derive conclusions based on observed differences, similarities, or relative outcomes.

- **Task Input:** A scientific passage containing information about different experimental conditions or groups.
- **Task Output:** The exact attribute value or Boolean answer that accurately reflects the comparison between different conditions.
- **Evaluation Metric:** Exact-match F1 score between the predicted and reference answers.

Input:

As the aforementioned studies have reported a requirement for or stimulus by trace elements or organic compounds for growth of Methanocaldococcus species in the presence of H2 and CO2, MGG medium was prepared without trace minerals and the following substances were added individually or in combination: yeast extract (0.1 g l^{-1}), selenate (0.05 g l^{-1}), tungstate (0.05 g l^{-1}), 1-fold trace mineral solution (Huber & Stetter, 2006; 10 ml l^{-1}) and 1-fold vitamin solution (Balch et al., 1979; 10 ml l^{-1}). Experiments in MGG medium without trace mineral solution resulted in slower growth and two- to fourfold lower final cell densities compared with the original culture medium. Adding selenate to the medium compensated for the effects on growth rate and final cell densities caused by leaving out the trace mineral solution, whereas the addition of tungstate and yeast extract had no influence on the doubling time.

Question:

"Is tungstate an essential growth component?"

Expected Answer:

```
<Answer>false</Answer>
```

D.4 Layout Structure and Semantic Region Recognition

D.4.1 Semantic Document Region Extraction

This task focuses on identifying and extracting key semantic units from scientific documents, such as author names, article titles, and journal names. The goal is to convert unstructured text into structured information to support document understanding and organization.

- **Task Input:** A passage containing various semantic units such as author names, titles, and other document-related metadata.
- **Task Output:** The exact semantic unit or attribute value extracted from the passage.
- **Evaluation Metric:** Exact-match F1 score between the predicted and reference answers.

Input:

"International Journal of Systematic and Evolutionary Microbiology (2011), 61, 1239–1245 DOI 10.1099/ijs.0.023663-0 Correspondence Annett Bellack annett.bellack@biologie.uni-regensburg.de Methanocaldococcus villosus sp. nov., a heavily flagellated archaeon that adheres to surfaces and forms cell-cell contacts Annett Bellack, Harald Huber, Reinhard Rachel, Gerhard Wanner and Reinhard Wirth Lehrstuhl fuer Mikrobiologie und Archaeenzentrum, Universitaet Regensburg, Universitaetsstrasse 31, 93053 Regensburg, Germany Zentrum fuer Elektronenmikroskopie der NWFIII, Universitaet Regensburg, Universitaetsstrasse 31, 93053 Regensburg, Germany Biozentrum der LMU, Department Biologie I, Großhadernerstrasse 4, 82152 Planegg"

Question:

"What is the name of the journal where this paper was published?"

Expected Answer:

```
<Answer>International
Journal of Systematic
and Evolutionary
Microbiology</Answer>
```

E JSON Example

In this section, we introduce the JSON format used for the MicrobeCollect IR workflow and for MicrobeQuest QA test cases.

E.1 JSON Structure of MicrobeCollect

This JSON format is generated by the MicrobeCollect agentic processing pipeline and further refined through cross-modal alignment. It captures extracted attribute values, supporting evidence, document anchors (e.g., OCR locations), and confidence scores.

value: The extracted attribute value. (e.g., `"0.5 M"`)

source_text: The exact text span from the document that supports the extracted value. (e.g., `"NaCl concentration was 0.5 M for optimal growth."`)

location: The OCR bounding box in pixel coordinates for the evidence text on the original page, typically formatted as `[x0, y0, x1, y1]`. (e.g., `[112, 235, 290, 260]`)

evidence_score: A model-assigned confidence score ranging from 0 to 100 that reflects the reliability of the extracted evidence. (e.g., `94`)

E.2 JSON Structure of MicrobeQuest

This JSON format is specifically designed for MicrobeQuest QA test cases, aimed at evaluating the performance of models.

case_id: A unique identifier for the test case (e.g., 'MB-S071-00001'),

version: The version number of the test case

timestamp: The timestamp when the test case was created or updated,
difficulty: The difficulty level of the test case (Easy, Medium, Hard),
task_category: The main category of the task,
task_subcategory: The subcategory of the task,
test_case:

- **input:**
  - **pdf_index:** The index of the relevant PDF document (e.g., 'S071'),
  - **image_index:** A list of image indices related to the test case,
  - **is_full_pdf:** A boolean indicating whether the entire PDF is available (e.g., true)
- **question:** The question to be answered in the test case (e.g., 'What is the source of the microorganism for the strain WAL1?'),
- **note:** The instructions for extracting the required information, including step-by-step guidelines,
- **few_shot_examples:** A list of example questions and expected answers to guide the extraction process,
- **expected_answer_type:** Specifies the type of expected answer (e.g., 'descriptive'),
- **expected_answer:** The expected output

F Model Performance

This section provides a brief overview of the model's performance across various benchmark tables and charts, covering accuracy score, F1 score, and BLEU score under different task types. It aims to offer a comprehensive evaluation of the model's strengths, weaknesses, and applicability across multiple metrics.

F.1 Tabular Results

We present three benchmark tables summarizing the model's performance in terms of accuracy score (Table 4), F1 score (Table 5), and BLEU score (Table 6).

F.2 Radar Chart Analysis

To provide a clearer understanding of the model's performance across different task categories, we present four radar charts 7, 8. Each chart corresponds to one major task type and illustrates the model's performance in terms of F1 score across its sub-tasks.

Task	Subtask	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13	M14	M15	M16	M17
Structured Information Extraction	Strain Entity Recognition and Normalization	0.300	0.432	0.148	0.276	0.000	0.263	0.362	0.330	0.490	0.267	0.537	0.315	0.417	0.451	0.444	0.478	0.480
	Strain Entity Resolution	0.550	0.506	0.774	0.727	0.286	0.562	0.588	0.570	0.660	0.500	0.550	0.498	0.503	0.511	0.712	0.541	0.600
	Strain Taxonomy Extraction	0.496	0.466	0.316	0.379	0.000	0.444	0.439	0.370	0.365	0.297	0.303	0.359	0.330	0.332	0.408	0.325	0.420
	Strain Physiological Characteristic Extraction	0.639	0.645	0.610	0.702	0.095	0.699	0.723	0.614	0.695	0.684	0.691	0.724	0.522	0.752	0.653	0.556	0.620
	Environmental Growth Parameter Extraction	0.562	0.690	0.504	0.498	0.000	0.629	0.622	0.605	0.644	0.599	0.644	0.526	0.432	0.574	0.667	0.488	0.460
	Strain Attribute Semantic Categorization	0.636	0.649	0.478	0.536	0.000	0.675	0.623	0.567	0.543	0.740	0.701	0.687	0.424	0.731	0.852	0.500	0.720
	Strain Culture Medium and Growth Condition Extraction	0.754	0.677	0.481	0.518	0.357	0.723	0.661	0.787	0.773	0.600	0.850	0.559	0.581	0.633	0.638	0.580	0.520
Multimodal Understanding	Table-based Strain Attribute Extraction	0.437	0.562	0.450	0.212	0.000	0.547	0.583	0.478	0.524	0.475	0.577	0.381	0.195	0.474	0.694	0.286	0.540
	Figure-based Strain Attribute Extraction	0.360	0.380	0.326	0.250	0.071	0.340	0.420	0.312	0.354	0.380	0.420	0.370	0.241	0.413	0.348	0.222	0.380
	Multimodal Strain Attribute Reasoning	0.600	0.557	0.500	0.600	0.000	0.613	0.625	0.408	0.550	0.611	0.625	0.396	0.620	0.590	0.673	0.740	0.540
Complex Semantic Reasoning	Multi-Entity Attribute Association	0.450	0.463	0.612	0.380	0.020	0.550	0.537	0.536	0.500	0.522	0.537	0.535	0.388	0.578	0.673	0.267	0.640
	Multi-value Priority Resolution	0.463	0.455	0.335	0.244	0.061	0.338	0.400	0.345	0.343	0.411	0.525	0.409	0.245	0.429	0.449	0.316	0.480
	Negation and Contrast Relationship Parsing	0.675	0.747	0.553	0.750	0.095	0.734	0.744	0.700	0.656	0.744	0.800	0.705	0.656	0.727	0.775	0.655	0.720
	Logical Condition Reasoning	0.580	0.784	0.512	0.520	0.000	0.625	0.775	0.436	0.579	0.711	0.784	0.634	0.714	0.690	0.640	0.680	0.560
	Cross-paragraph Entity Tracking	0.435	0.544	0.413	0.391	0.020	0.438	0.468	0.405	0.378	0.379	0.437	0.431	0.253	0.422	0.480	0.400	0.420
	Implicit Conclusion Generation	0.444	0.468	0.388	0.425	0.000	0.395	0.446	0.421	0.407	0.411	0.456	0.495	0.361	0.497	0.388	0.464	0.540
	Multi-instance Comparative Reasoning	0.425	0.426	0.257	0.202	0.000	0.388	0.475	0.543	0.500	0.400	0.475	0.294	0.287	0.330	0.250	0.321	0.260
Layout Structure and Semantic Region Recognition	Semantic Document Region Extraction	0.662	0.280	0.506	0.288	0.000	0.675	0.375	0.710	0.250	0.667	0.250	0.753	0.282	0.731	0.709	0.282	0.680
Overall Accuracy Score		0.526	0.541	0.454	0.439	0.056	0.535	0.548	0.508	0.512	0.522	0.565	0.504	0.414	0.548	0.581	0.450	0.532

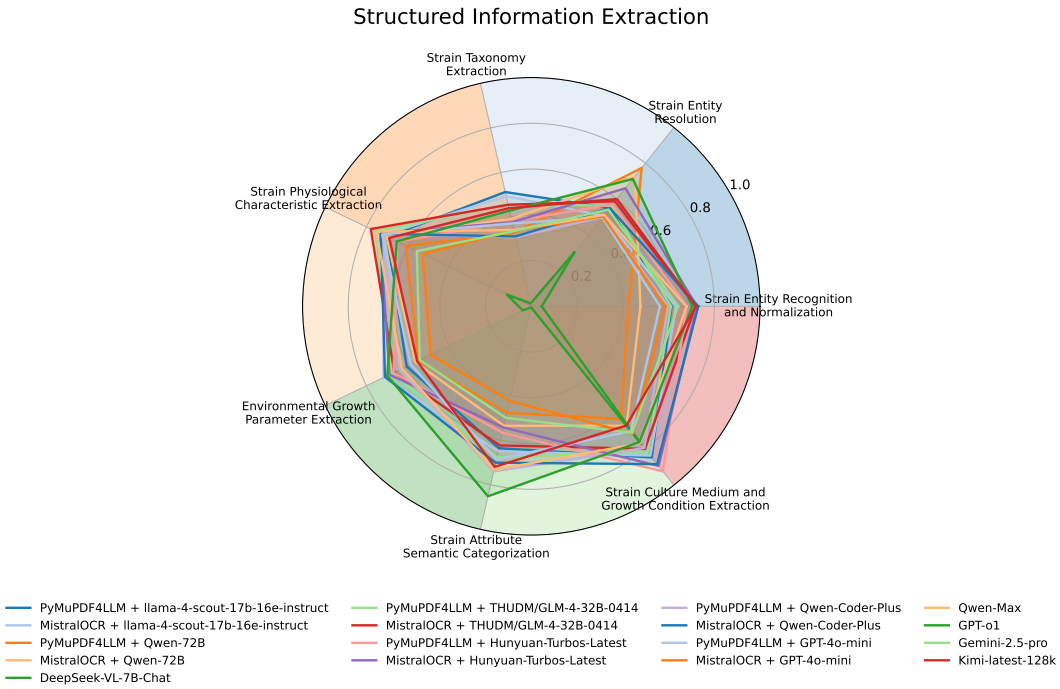
Table 4: Accuracy performance of all models on the MicrobeQuest multimodal benchmark. Blue text indicates open-source models, orange text signifies closed-source models and red number indicates the best-performances model. Model identifiers:
M1 = PyMuPDF4LLM + llama-4-scout-17b-16e-instruct,
M2 = MistralOCR + llama-4-scout-17b-16e-instruct,
M3 = PyMuPDF4LLM + Qwen-72B,
M4 = MistralOCR + Qwen-72B,
M5 = DeepSeek-VL-7B-Chat,
M6 = PyMuPDF4LLM + THUDM/GLM-4-32B-0414,
M7 = MistralOCR + THUDM/GLM-4-32B-0414,
M8 = PyMuPDF4LLM + Hunyuan-Turbos-Latest,
M9 = MistralOCR + Hunyuan-Turbos-Latest,
M10 = PyMuPDF4LLM + Qwen-Coder-Plus,
M11 = MistralOCR + Qwen-Coder-Plus,
M12 = PyMuPDF4LLM + GPT-4o-mini,
M13 = MistralOCR + GPT-4o-mini,
M14 = Qwen-Max,
M15 = GPT-o1,
M16 = Gemini-2.5-pro,
M17 = Kimi-latest-128k.

Task	Subtask	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13	M14	M15	M16	M17
Structured Information Extraction	Strain Entity Recognition and Normalization	0.613	0.633	0.425	0.477	0.046	0.610	0.718	0.679	0.724	0.598	0.730	0.562	0.586	0.678	0.704	0.642	0.720
	Strain Entity Resolution	0.550	0.506	0.774	0.727	0.303	0.562	0.588	0.570	0.660	0.500	0.550	0.498	0.503	0.511	0.712	0.541	0.600
	Strain Taxonomy Extraction	0.512	0.494	0.331	0.393	0.013	0.458	0.455	0.381	0.377	0.309	0.313	0.372	0.340	0.342	0.429	0.338	0.440
	Strain Physiological Characteristic Extraction	0.700	0.719	0.610	0.702	0.119	0.757	0.779	0.650	0.734	0.732	0.730	0.724	0.527	0.752	0.653	0.556	0.689
	Environmental Growth Parameter Extraction	0.605	0.717	0.548	0.563	0.041	0.676	0.665	0.668	0.688	0.638	0.710	0.585	0.488	0.628	0.699	0.541	0.554
	Strain Attribute Semantic Categorization	0.636	0.650	0.478	0.536	0.005	0.676	0.624	0.567	0.544	0.740	0.701	0.687	0.424	0.734	0.852	0.500	0.720
	Strain Culture Medium and Growth Condition Extraction	0.845	0.835	0.629	0.668	0.770	0.816	0.797	0.923	0.893	0.792	0.882	0.698	0.713	0.766	0.759	0.707	0.667
Multimodal Understanding	Table-based Strain Attribute Extraction	0.483	0.605	0.482	0.247	0.029	0.581	0.622	0.513	0.558	0.519	0.615	0.425	0.216	0.516	0.694	0.415	0.548
	Figure-based Strain Attribute Extraction	0.394	0.414	0.388	0.365	0.081	0.375	0.457	0.356	0.397	0.383	0.426	0.415	0.276	0.462	0.388	0.413	0.410
	Multimodal Strain Attribute Reasoning	0.627	0.635	0.526	0.632	0.022	0.655	0.656	0.427	0.573	0.619	0.644	0.417	0.642	0.605	0.689	0.740	0.559
Complex Semantic Reasoning	Multi-Entity Attribute Association	0.477	0.485	0.630	0.416	0.062	0.607	0.568	0.571	0.513	0.546	0.559	0.565	0.410	0.598	0.685	0.301	0.651
	Multi-value Priority Resolution	0.660	0.676	0.485	0.535	0.112	0.597	0.635	0.526	0.547	0.613	0.703	0.592	0.543	0.616	0.676	0.518	0.695
	Negation and Contrast Relationship Parsing	0.813	0.844	0.607	0.786	0.116	0.847	0.861	0.795	0.768	0.854	0.893	0.838	0.727	0.852	0.870	0.767	0.836
	Logical Condition Reasoning	0.616	0.805	0.541	0.548	0.017	0.645	0.800	0.458	0.593	0.728	0.811	0.663	0.726	0.722	0.675	0.711	0.610
	Cross-paragraph Entity Tracking	0.535	0.637	0.493	0.519	0.111	0.528	0.561	0.497	0.469	0.454	0.515	0.535	0.346	0.511	0.563	0.469	0.507
	Implicit Conclusion Generation	0.552	0.577	0.437	0.497	0.026	0.485	0.512	0.490	0.483	0.516	0.529	0.585	0.428	0.592	0.483	0.580	0.594
	Multi-instance Comparative Reasoning	0.536	0.534	0.377	0.350	0.011	0.500	0.552	0.609	0.586	0.494	0.558	0.423	0.429	0.446	0.455	0.417	0.422
Layout Structure and Semantic Region Recognition	Semantic Document Region Extraction	0.936	0.561	0.759	0.558	0.085	0.906	0.639	0.930	0.513	0.908	0.544	0.960	0.520	0.940	0.938	0.520	0.923
Overall F1 Score		0.616	0.629	0.529	0.529	0.109	0.627	0.638	0.589	0.590	0.608	0.634	0.586	0.491	0.626	0.662	0.538	0.619

Table 5: F1 performance of all models on the MicrobeQuest multimodal benchmark. Blue text indicates open-source models, orange text signifies closed-source models and red number indicates the best-performances model. Model identifiers:
M1 = PyMuPDF4LLM + llama-4-scout-17b-16e-instruct,
M2 = MistralOCR + llama-4-scout-17b-16e-instruct,
M3 = PyMuPDF4LLM + Qwen-72B,
M4 = MistralOCR + Qwen-72B,
M5 = DeepSeek-VL-7B-Chat,
M6 = PyMuPDF4LLM + THUDM/GLM-4-32B-0414,
M7 = MistralOCR + THUDM/GLM-4-32B-0414,
M8 = PyMuPDF4LLM + Hunyuan-Turbos-Latest,
M9 = MistralOCR + Hunyuan-Turbos-Latest,
M10 = PyMuPDF4LLM + Qwen-Coder-Plus,
M11 = MistralOCR + Qwen-Coder-Plus,
M12 = PyMuPDF4LLM + GPT-4o-mini,
M13 = MistralOCR + GPT-4o-mini,
M14 = Qwen-Max,
M15 = GPT-o1,
M16 = Gemini-2.5-pro,
M17 = Kimi-latest-128k.

Task	Subtask	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13	M14	M15	M16	M17
Structured Information Extraction	Strain Entity Recognition and Normalization	0.232	0.287	0.141	0.195	0.008	0.259	0.334	0.285	0.372	0.240	0.369	0.215	0.265	0.264	0.286	0.249	0.261
	Strain Entity Resolution	0.098	0.090	0.138	0.129	0.054	0.100	0.104	0.101	0.117	0.089	0.098	0.088	0.089	0.091	0.127	0.096	0.107
	Strain Taxonomy Extraction	0.111	0.108	0.076	0.089	0.001	0.100	0.099	0.090	0.089	0.072	0.075	0.084	0.078	0.079	0.096	0.079	0.092
	Strain Physiological Characteristic Extraction	0.171	0.175	0.135	0.142	0.019	0.193	0.192	0.158	0.179	0.184	0.186	0.151	0.100	0.162	0.139	0.099	0.167
	Environmental Growth Parameter Extraction	0.165	0.197	0.149	0.146	0.004	0.181	0.183	0.185	0.191	0.174	0.197	0.156	0.127	0.168	0.185	0.152	0.151
	Strain Attribute Semantic Categorization	0.174	0.180	0.128	0.144	0.001	0.174	0.163	0.144	0.141	0.198	0.187	0.184	0.121	0.196	0.228	0.123	0.181
	Strain Culture Medium and Growth Condition Extraction	0.284	0.279	0.220	0.227	0.130	0.259	0.240	0.286	0.233	0.299	0.186	0.243	0.250	0.287	0.263	0.397	0.229
Multimodal Understanding	Table-based Strain Attribute Extraction	0.103	0.128	0.112	0.058	0.003	0.122	0.133	0.122	0.129	0.110	0.137	0.089	0.068	0.104	0.132	0.164	0.113
	Figure-based Strain Attribute Extraction	0.070	0.073	0.074	0.071	0.014	0.073	0.090	0.076	0.089	0.094	0.089	0.073	0.082	0.086	0.069	0.075	0.075
	Multimodal Strain Attribute Reasoning	0.109	0.114	0.097	0.114	0.003	0.116	0.117	0.077	0.105	0.111	0.113	0.076	0.120	0.110	0.134	0.132	0.112
Complex Semantic Reasoning	Multi-Entity Attribute Association	0.144	0.139	0.193	0.131	0.008	0.172	0.166	0.171	0.163	0.159	0.164	0.166	0.130	0.180	0.209	0.095	0.173
	Multi-value Priority Resolution	0.140	0.137	0.090	0.109	0.017	0.112	0.130	0.104	0.122	0.132	0.158	0.122	0.118	0.127	0.149	0.093	0.134
	Negation and Contrast Relationship Parsing	0.145	0.156	0.111	0.139	0.018	0.153	0.155	0.143	0.141	0.158	0.157	0.153	0.134	0.157	0.154	0.140	0.163
	Logical Condition Reasoning	0.125	0.160	0.100	0.114	0.004	0.131	0.160	0.101	0.126	0.147	0.169	0.126	0.159	0.137	0.120	0.151	0.109
	Cross-paragraph Entity Tracking	0.164	0.190	0.150	0.162	0.013	0.155	0.164	0.162	0.154	0.136	0.154	0.156	0.106	0.156	0.163	0.144	0.138
	Implicit Conclusion Generation	0.122	0.136	0.102	0.122	0.003	0.107	0.131	0.123	0.119	0.122	0.119	0.127	0.101	0.129	0.104	0.150	0.128
	Multi-instance Comparative Reasoning	0.108	0.106	0.079	0.075	0.001	0.106	0.116	0.129	0.124	0.103	0.116	0.100	0.104	0.116	0.104	0.100	0.105
Layout Structure and Semantic Region Recognition	Semantic Document Region Extraction	0.636	0.422	0.433	0.380	0.022	0.623	0.488	0.623	0.410	0.596	0.424	0.658	0.398	0.645	0.636	0.398	0.605
Overall BLEU Score		0.172	0.171	0.140	0.142	0.018	0.174	0.176	0.171	0.167	0.174	0.172	0.165	0.142	0.177	0.183	0.158	0.169

Table 6: BLEU performance of all models on the MicrobeQuest multimodal benchmark. Blue text indicates open-source models, orange text signifies closed-source models and red number indicates the best-performances model. Model identifiers:
M1 = PyMuPDF4LLM + llama-4-scout-17b-16e-instruct,
M2 = MistralOCR + llama-4-scout-17b-16e-instruct,
M3 = PyMuPDF4LLM + Qwen-72B,
M4 = MistralOCR + Qwen-72B,
M5 = DeepSeek-VL-7B-Chat,
M6 = PyMuPDF4LLM + THUDM/GLM-4-32B-0414,
M7 = MistralOCR + THUDM/GLM-4-32B-0414,
M8 = PyMuPDF4LLM + Hunyuan-Turbos-Latest,
M9 = MistralOCR + Hunyuan-Turbos-Latest,
M10 = PyMuPDF4LLM + Qwen-Coder-Plus,
M11 = MistralOCR + Qwen-Coder-Plus,
M12 = PyMuPDF4LLM + GPT-4o-mini,
M13 = MistralOCR + GPT-4o-mini,
M14 = Qwen-Max,
M15 = GPT-o1,
M16 = Gemini-2.5-pro,
M17 = Kimi-latest-128k.



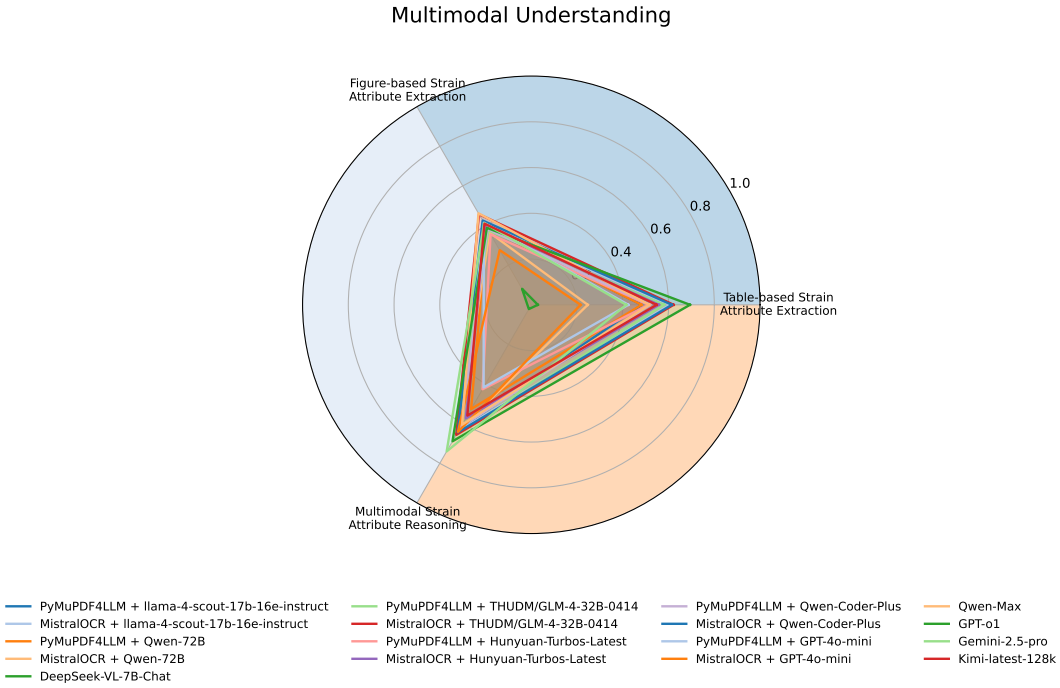
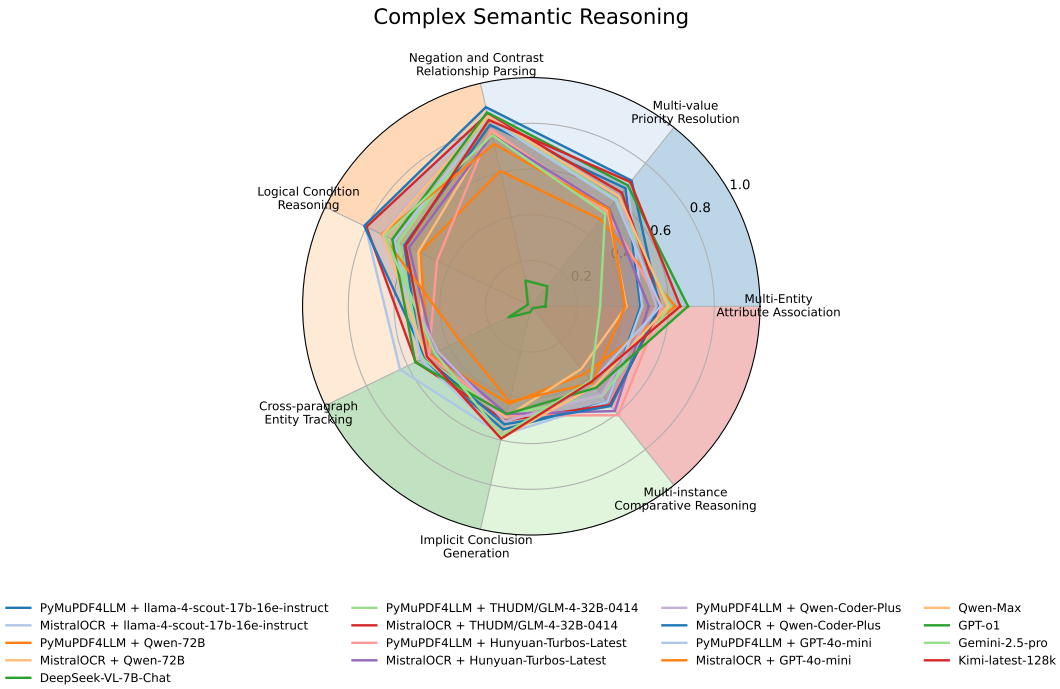


Figure 7: Radar charts showing F1 scores across sub-tasks in Structured Information Extraction and Multimodal Understanding.



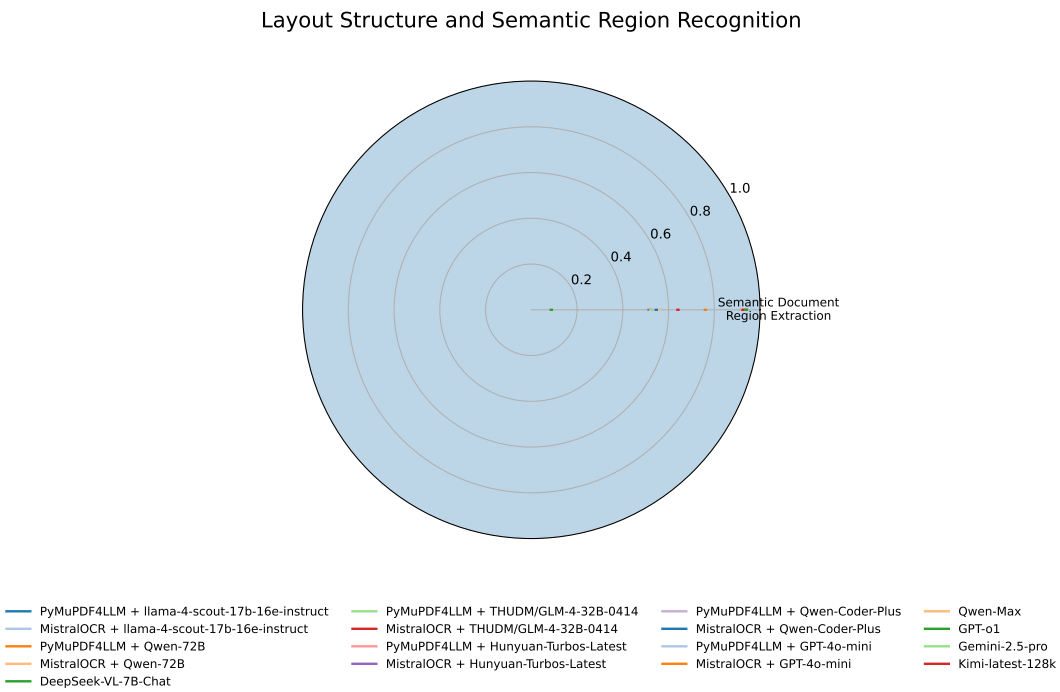


Figure 8: Radar charts showing F1 scores across sub-tasks Complex Semantic Reasoning and Layout and Semantic Region Recognition.

G Error Analysis

We categorize the representative error types encountered in our tasks and provide illustrative examples in the following tables.

- **Domain Knowledge Dependency**: Errors that occur when interpreting specialized microbiological terminology or context require specific domain expertise. Respective examples are seen in Table 7.
- **Multi-Value Confusion**: Errors caused by multiple plausible values within the same context, leading to ambiguity in extraction. Respective examples are seen in Table 8.
- **Multi-Entity Confusion**: Errors arising from the presence of multiple, co-referenced, or closely located entities, which complicate precise entity boundary detection. Respective examples are seen in Table 9.
- **Syntactic Complexity**: Errors resulting from complex sentence structures or long-range dependencies that impede accurate parsing and understanding. Respective examples are seen in Table 10.
- **Structural Parsing Challenges**: Errors associated with correctly identifying relationships presented in tables, graphs, or other structured document elements. Respective examples are seen in Table 11.

H MicrobeCollect Platform

To support annotation for microbial IR tasks, we developed MicrobeCollect, an expert-in-the-loop platform that integrates PDF visualization, model output review, and efficient manual correction. Figure 9 shows a demonstration using content from Yu and Leadbetter (2020), which is used here for illustrative purposes only.
In the interface, the left panel displays the original PDF, while the right panel shows extracted fields. When a field is selected, its source is highlighted in the document. In this example, a human annotator is correcting an error in the model's output.

Question	Context	Expected Answer	Predicted Answer	Error Type
What is the full name of a strain (species + strain)?	Based on the results from phylogenetic, morphological, and protein analyses, we conclude that the novel strain represents a new species within the genus Methanocaldococcus, for which the name Methanocaldococcus villosus sp. nov. is proposed (type strain KIN24-T80 = DSM 22612 = JCM 16315.	Methanocaldococcus villosus KIN24-T80	Methanocaldococcus villosus sp.	Domain knowledge dependency
Is strain Ivanov the same as DSM 2611?	One novel isolate, *Methanobacterium* sp. strain Ivanov, was grown on H_2-CO_2, and the stable-carbon isotopic fractionations that occurred during the synthesis of methane, cell carbon, and lipids were determined. These results were used to examine the anomalous relationship between the isotopic and chemical compositions of natural gas in deep subsurface oil field environments.	true	false	Domain knowledge dependency
What is the class of the microorganism for the strain JAL-1?	*Mycobacterium tuberculosis* is classified within the class Actinomycetia, a group of high G+C Gram-positive bacteria.	Actinomycetia	Methanococci	Domain knowledge dependency

Table 7: Representative Error Cases Involving Domain Knowledge Dependency.

Question	Context	Expected Answer	Predicted Answer	Error Type
In what specific habitat or environment does this organism typically grow for the strain 6A8?	A novel acidophilic, hydrogenotrophic methanogen, designated strain $6A8^T$, was isolated from an acidic (pH 4.0–4.5), ombrotrophic (rain-fed) bog near Ithaca, NY, USA. Cultures were dimorphic, containing thin rods (0.2–0.3 μm diameter, 0.8–3.0 μm length) and irregular cocci (0.2–0.8 μm diameter).	Water/sediments	Soil	Multi-Value Confusion
In what environment was the strain M7 isolated?	A chimney sample was collected from the 13 N hydrothermal field during the Hero cruise (1991), on the East Pacific Rise at a depth of 2600 m.; The new strain was isolated from a chimney sample collected from the 13 °N hydrothermal field (12,48 °N, 103,56°W) during the 'Hero' cruise (1991), on the East Pacific Rise at a depth of 2600 m.	Deep-sea hydrothermal chimney sample collected on the East Pacific Rise (13°N; 103°W) at a depth of 2600 m	A chimney sample collected from the 13°N hydrothermal field during the Hero cruise (1991), on the East Pacific Rise at a depth of 2600 m	Multi-Value Confusion
What substrates do not support the organism's growth for the strain M7?	The new isolate utilized methanol in addition to methylamines but not H_2:CO_2, formate, or acetate. The optimal initial pH for growth is 7.7. Trimethylamine, dimethylamine, methylamine, and methanol are substrates for growth and methanogenesis. In contrast, acetate, formate, ethanol, dimethyl sulfide, acetone, 2-butanol, 2-propanol, 1-propanol, and H_2:CO_2 do not support growth. No growth factors are required, but yeast extract and trypticase greatly stimulate growth.	Acetate, formate, ethanol, dimethyl sulfide, acetone, 2-butanol, 2-propanol, 1-propanol, and H_2:CO_2	H_2:CO_2, formate, or acetate	Multi-Value Confusion

Table 8: Representative Error Cases Involving Multi-Value Confusion.

Question	Context	Expected Answer	Predicted Answer	Error Type
What is the Gram reaction for the strain kuznetsov?	Strain kuznetsov stained Gram-negative, whereas both strains Ivanov and Omeliansky stained Gram-positive.	Gram-negative	Gram-positive	Multi-Entity confusion
What is the lower bound of the optimal growth temperature range for the strain AK-7?	Methanogenium boonei (boone.i. N.L. gen. n. boonei of Boone; named in honor of David R. Boone, who has made many contributions to the ecology, physiology, and taxonomy of methanogens). The organism takes the form of irregular cocci 1.0 to 2.5 μm in diameter, occurring singly, and is nonmotile. CO_2 plus H_2 or formate serves as the sole catabolic substrate, with methane as the end product. The fastest growth occurred at 19.4, with a salinity of 0.3 to 0.5 M Na+ and a pH of 6.4 to 7.8. It was isolated from permanently cold, anoxic marine sediments at Skan Bay, Alaska. The type strain is AK-7 (OCM 787/DSMZ 17338).	19.4	15	Multi-Entity confusion
What is the G+C content for the strain AK-3?	The Tm values for strains AK-7, AK-3, and AK-8 were 85.7, 83.2, and 84.2, respectively, which correspond to DNA G+C contents of 49.7, 43.6, and 46.2 mol%, respectively.	43.6	49.7	Multi-Entity confusion

Table 9: Representative Error Cases Involving Multi-Entity Confusion.

Question	Context	Expected Answer	Predicted Answer	Error Type
Which elements or compounds do not stimulate the organism's growth for the strain SD-1?	Strain SD-1 grew in medium containing trimethylamine, dimethylamine, monomethylamine, or methanol as the catabolic substrate. No growth occurred in medium supplemented with 50 mM acetate, 100 kPa of H, plus 20 kPa of CO,, or 5 mM dimethyl sulfide as the catabolic substrate. When 100 kPa of H, plus trimethylamine was added as the catabolic substrate, cultures formed the same quantity of methane and grew at the same rate as controls in medium containing only trimethylamine. With trimethylamine as the catabolic substrate, the cells grew fastest at 42°C (Fig. l), at pH 7.8 (Fig. 2), and in the presence of 0.9 to 3.5 M Na+ Fig.3). Cells grew rapidly (specific growth rate, 0.015 h-) in mineral medium with no organic compound other than trimethylamine added, vitamins did not stimulate growth.	Vitamins	not provided	Syntactic Complexity
How many solutes are there in the culture medium?	YPS MEDIUM: Sea salts (Sigma) 35.00 g, PIPES 3.46 g, Yeast extract 1.00 g, Peptone 4.00 g, Elemental sulphur 5.00 g, NH_4Cl 0.50 g, KH_2PO_4 0.35 g, $CaCl_2$ 0.20 g, $FeCl_3$ 6.70 mg, Na_2WO_4 2.90 mg, Resazurin 0.10 mg, $Na_2S{\cdot}9H_2O$ 0.25 g, plus instructions on pH adjustment, nitrogen flushing, and sterilization.	12	not provided	Syntactic Complexity
Is the strain anaerobic for the strain C?	Strain CT, a non-motile, mesophilic, hydrogenotrophic, methanogenic bacterium, was isolated from an anaerobic digester used for the treatment of raw cassava-peel waste in Congo. The cells were rods, 0.4–0.5 × 2–10 µm in size, and stained Gram-positive. Hydrogen and carbon dioxide were the only substrates that supported growth and methane production. Methane production, but not growth, occurred with CO_2 in the presence of either 2-propanol, 2-butanol or cyclopentanol as hydrogen donors. The temperature range for growth was 25–50 °C, the optimum being between 37 and 42 °C.	true	false	Syntactic Complexity

Table 10: Representative Error Cases Involving Syntactic Complexity.

Question	Context	Expected Answer	Predicted Answer	Error Type
What is the upper bound of the optimal growth temperature range for the strain Sar?	the three rods, as Microorganism Culture condition ARNr 16S phylogeny Optimum (and range) Substrate for growth. Most closely related species (similarity %) Temperature (°C) Salinity (g/l) pH H_2/CO_2, Formate Methanol Acetate Alcohols. Methanobacterium bryantii 37–39 n.d. 6.9–7.2 + - - - +. Methanobacterium formicicum? 37–45 n.d. 6.6–7.8 + + - - +. /RiH2 (Camargue) ○ ○ ○ 2B Mb. bryantii 99.2-FCam (Camargue)* ○ ○ 2B Mb. formicicum 97.9 'FPi (Pila) ○ ○ ○ *Mb. bryantii* 96.5 Methanosarcina barkeri ○ - + + - Methanosarcina mazei? - + + n.d. 'Sar (Camargue)! ○ ○ ○ Ms. barkeri 99.0 'SarPi (Pila) ○ 6 ○ 8 Ms. mazei 99.8 Methanoculleus marisnigri? + - + CoCam ○ ○ Me. marisnigri 98.4 (Camargue)*. *Characteristics of related species according to Garcia [8]. Reference strains. Soil of origin. Concentrations $>$ 60 g/l of NaCl not tested. n.d.: not determined.	37	not provided	Structural Parsing Challenges
What is the optimal growth temperature for the strain omeliansky?	694 BELYAEV ET AL. SPECIFIC GROWTH RATE (hr^{-1}) 10 20 30 40 50 60 TEMPERATURE (°C) FIG. 2. Relationship between growth rate and incubation temperature for oil field Methanobacterium sp. Symbols: O, strain ivanov; A, strain kuznetsov; O, strain omeliansky. Three strains at their respective growth temperature optima was 16 to 18 h. The pH optimum for growth was between 6.5 to 7.2 for strain omeliansky and 7.0 to 7.4 for the other strains. Strain omeliansky grew within the pH range of 6.0 to 7.4, whereas strains ivanov and kuznetsov grew within the pH range of 6.5 to 8.2. Nutritional studies were performed in maintenance medium at the optimum temperature for growth of each strain.	37	not provided	Structural Parsing Challenges
What is the maximum length of Cell morphology for the strain NOBI-1?	ND, No data available. Numbers in parentheses for the optimum temperature, pH and NaCl concentration indicate the range allowing growth. Characteristic 1 2 3 4 5 6 7 8 Cell morphology Rod Rod* Rod Angular, Highly irregular – Irregular Irregular – Sheathed crystal-like plate cocci cocci cocci rod or disc-shaped Cell width (µm) 0.7–1.0 0.2–0.3 0.7 1.5 1.0–3.0 1–2 1–2.5 0.5 Cell length (µm) 2–8+ 0.8–3.0 1.5–2.0 1.6–2.8 1.0–3.0 1–2 1–2.5 7.4–$>$100 G+C content 56.3 ND 48.84 47.58 51.6 59.4 54.8 45 (mol%) Optimum temperature (°C): 50 (35–55) 37 (10–40) 40 (30–45) 40 (17–41) 20–25 37 (25–55) 37–40 (25–45) 30–37 Optimum pH: 7 (6.7–8.0), 5 (4.0–5.8), 6.1–6.9 (5.9–7.7), 6.5–7.5, 6.8–7.3 (6.0–8.3), 6.7 (5.5–8.0), 7 (6.6–8.8), 6.6–7.4 Optimum NaCl (g/l): 0 (0–15), ND, ND, 10 (4–54), 26.9, $<$10, 8–12 (0–70), ND Motility: – ND + + – – + + Substrate utilization: Formate – – – – – – – – –(2-Propanol/CO_2): – – ND – – + + – Growth requirements: Yeast extract + + + – + – + ND Acetate + + + + + + + ND *Cells sometimes become spherical with a diameter of 0.3–0.8 µm. Determined by buoyant density. Determined by thermal denaturation. Rod-shaped cells with blunt-ends. Often form multicellular filaments. Methanogenic and strictly anaerobic members of the order Methanomicrobiales, phylum Euryarchaeota, domain Archaea. Can use H_2 or formate for growth and methane production. The type species is Methanolinea tarda.	8	not provided	Structural Parsing Challenges

Table 11: Representative Error Cases Involving Structural Parsing Challenges.

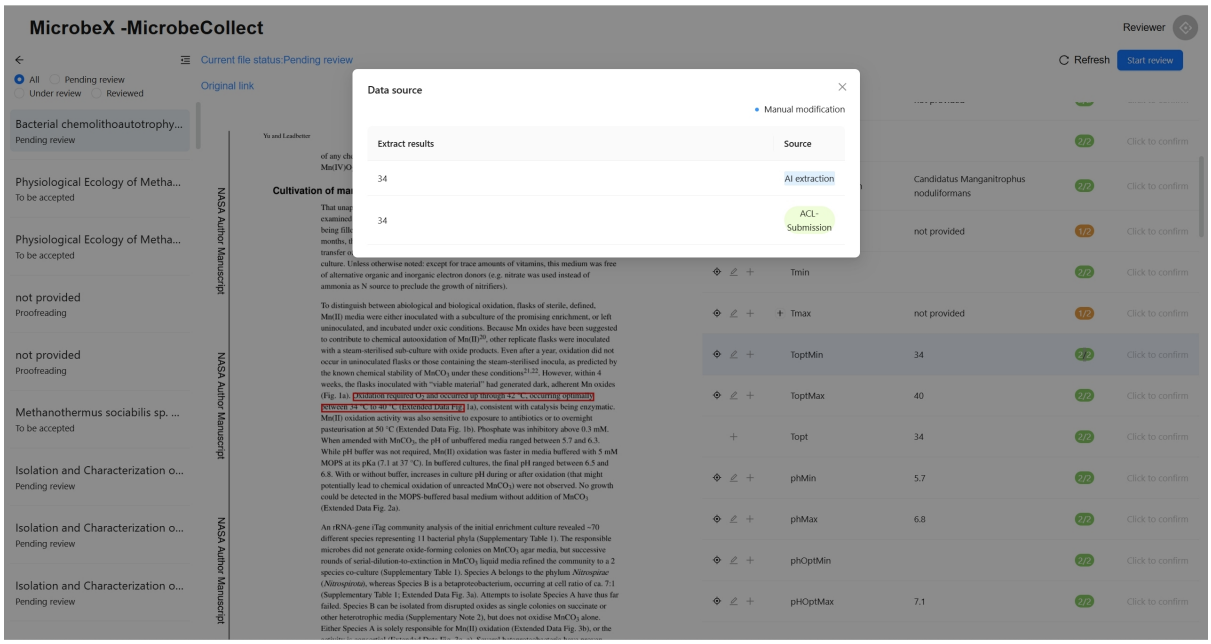


Figure 9: This example from the MicrobeCollect platform showcases the 'Microbiology Expert Final Review Stage' (Stage 3). The highlighted red box indicates where the data was extracted from. And the final review group compares both human validation and AI output results on this page.