PolyReal: A Benchmark for Real-World Polymer Science Workflows

Supplementary Material

6. Difficulty-Graded Evaluation Results

To provide a more granular understanding of MLLM capabilities, we stratified the PolyReal dataset into three difficulty levels: Easy (280 samples), Medium (165 samples), and Hard (100 samples). This stratification allows us to distinguish between superficial perception capabilities and deep, expert-level scientific reasoning.

6.1. Difficulty Grading Criteria

To ensure a fine-grained assessment of MLLM capabilities, our domain expert team manually categorized the 545 samples into three difficulty tiers, including Easy, Medium, and Hard. This classification is based on the complexity of information retrieval and the depth of scientific reasoning required.

Easy (51.4%): Direct Recognition and Single-Step Application. This category encompasses tasks that rely on direct visual recognition or single-step knowledge retrieval. These questions test whether a model possesses the fundamental "vocabulary" of polymer science and can apply basic principles to clear, unambiguous visual stimuli. A representative example is the Foundational Knowledge Application module (as shown in Figure 4, Case 1). Here, the reasoning process is straightforward: once the model comprehends the fundamental definition of a contact angle, it can directly recognize the visual feature (i.e., the magnitude of the angle) and perform the ranking task immediately without requiring complex intermediate logic or derivation.

Medium (30.3%): Multi-Step Reasoning and Cross-Modal Alignment. The medium-tier tasks elevate the challenge by requiring multi-step reasoning, cross-modal alignment, and specific domain application. Unlike the "Easy" tasks, the answer cannot be retrieved directly; it must be deduced through a chain of thought that bridges visual cues with domain-specific knowledge. This level is best exemplified by the Experiment Mechanism Reasoning module (see Figure 1). In these scenarios, the model must parse a professional reaction diagram, align the text annotations with chemical structures, and reason through the reaction pathway to infer intermediate steps or catalytic cycles.

Hard (18.3%): Holistic Synthesis and Multidisciplinary Analysis. These "Challenge Problems" represent the bottleneck of current AI systems, requiring expert-level synthesis to interpret highly unstructured, noisy, or abstract data within complex contexts. This category primarily includes Raw Data Extraction, Performance and Application Exploration, and notably, Lab Safety Analysis. For instance, in safety analysis tasks, the model must navigate cluttered, real-world laboratory environments. It requires a holistic, multidisciplinary understanding (integrating chemical properties, physical principles such as airflow dynamics, and safety protocols) to identify context-dependent hazards that are not explicitly labeled. Similarly, interpreting raw spectra requires the model to rigorously filter signal from noise, handling severe data ambiguity that typically stumps generalist models.

6.2. Data Distribution and Difficulty Analysis

We analyze the effectiveness of the dataset distribution through the lens of model performance reported in Table 6. The experimental results empirically validate the distinctiveness of the three difficulty tiers.

Validation of Difficulty Stratification. The results in Table 6 demonstrate a consistent monotonic degradation in performance across all 15 evaluated models as task difficulty increases, confirming that the stratification criteria effectively map to model capabilities. For instance, the state-of-the-art model O3 exhibits a clear step-wise decline in F1-score, dropping from 0.686 on the Easy subset to 0.561 on Medium, and finally to 0.430 on Hard. This trend indicates that the "Hard" subset successfully isolates complex reasoning and holistic synthesis from simpler perceptual tasks(As shown in Figure 6). Furthermore, the "Hard" category acts as a critical discriminator between closed-source and open-source models. While top-tier proprietary models maintain viable performance on these complex tasks, many open-weights models experience a severe capability collapse (e.g., DeepSeek-VL2 drops to an F1-score of 0.096), highlighting that the benchmark possesses a high ceiling necessary for evaluating future agents with strong reasoning abilities.

Disparity Between Recall and Precision. A critical behavioral trend observed in Table 6 is the systematic divergence between Recall (R) and Precision (P). This phenomenon is predominantly manifested in the "Easy" tier, where models consistently achieve disproportionately high Recall scores. For instance, Gemini-2.5-Flash achieves a near-perfect Recall of 0.981, yet its Precision remains remarkably low at 0.464. This gap suggests a prevalence of "defensive verbosity" in model responses. When addressing domain-specific scientific queries, generalist MLLMs tend to generate extensive, exhaustive explanations to ensure all potential key points are covered. While this strategy successfully retrieves the correct information in simpler contexts (thereby boosting Recall), it inevitably introduces a substantial amount of redundant context or subtle factual hallucinations, which severely penalizes Precision. In prac-

Table 6. Performance comparison of different models across three difficulty levels (**Hard**, **Medium**, **Easy**) and the **Overall** weighted average. **P**, **R**, and **F1** represent Precision, Recall, and F1-score, respectively. The best results are in **bold** and the second best are underlined.

	Hard			Medium			Easy			Overall		
Model	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
Closed-source Models												
O3	0.458	0.503	**0.430**	0.544	0.672	**0.561**	0.621	0.853	**0.686**	0.567	0.733	**0.601**
GPT-5	0.432	**0.546**	0.419	0.483	0.692	0.520	0.590	0.876	0.669	0.529	0.760	0.578
Claude-Sonnet-4.5-Thinking	0.392	0.432	0.365	0.480	0.638	0.513	0.559	0.820	0.632	0.503	0.692	0.546
Gemini-2.5-Pro	0.416	0.455	0.378	0.465	0.679	0.512	0.519	0.853	0.608	0.483	0.726	0.536
Grok-4	0.464	0.341	0.338	0.567	0.524	0.495	0.669	0.676	0.625	0.600	0.568	0.533
Gemini-2.5-Flash	0.330	0.341	0.289	0.425	**0.716**	0.504	0.464	**0.981**	0.598	0.427	**0.783**	0.512
Gemini-2.0-Flash-Thinking	**0.477**	0.286	0.303	**0.598**	0.401	0.423	**0.696**	0.540	0.555	**0.625**	0.450	0.468
GPT-4o	0.469	0.180	0.219	0.582	0.298	0.345	0.636	0.396	0.436	0.587	0.323	0.367
GPT-4o-mini	0.412	0.175	0.207	0.497	0.244	0.283	0.537	0.317	0.343	0.501	0.268	0.299
Open-source Models												
Qwen3-VL-235B-A22B-Thinking	0.443	0.390	**0.359**	0.527	**0.558**	**0.498**	**0.620**	**0.729**	**0.628**	0.558	**0.615**	**0.538**
Qwen3-VL-32B-Thinking	0.408	**0.405**	0.355	0.509	0.550	0.490	0.577	0.723	0.608	0.525	0.612	0.525
DeepSeek-R1	0.366	0.393	0.339	0.464	0.548	0.462	0.503	0.705	0.549	0.466	0.600	0.484
Intern-S1	0.429	0.258	0.262	0.531	0.433	0.400	0.599	0.573	0.499	0.548	0.473	0.427
Qwen2.5-VL-72B-Instruct	**0.436**	0.190	0.222	**0.564**	0.270	0.318	0.600	0.349	0.390	**0.559**	0.295	0.337
DeepSeek-VL2	0.291	0.077	0.096	0.329	0.095	0.117	0.384	0.127	0.153	0.350	0.108	0.132

tical polymer science workflows, this behavior is suboptimal, as it forces researchers to manually filter relevant insights from noisy outputs.

Conversely, the performance dynamics shift dramatically within the "Hard" difficulty tier, where we observe a precipitous decline in Recall alongside low Precision. As shown in Table 6, even capable models like GPT-4o see their Recall drop to 0.180 on Hard tasks, while open-source models such as DeepSeek-VL2 fall to a Recall of just 0.077. This "Recall Collapse" indicates that the strategy of verbose generation becomes ineffective when models face complex, multidisciplinary challenges (e.g., holistic synthesis or mechanism reasoning). In these scenarios, models fail to identify the core scientific principles entirely, resulting in the omission of essential Key Points rather than merely burying them in noise. This distinction strongly validates the robustness of our expert-defined difficulty stratification: the "Hard" subset successfully exposes the absolute reasoning boundaries of current MLLMs, confirming that future improvements must focus on deepening genuine scientific reasoning capabilities rather than simply optimizing for conversational comprehensiveness.

7. Implementation and Evaluation Details

This section outlines the comprehensive protocols ensuring the reproducibility and rigor of the *PolyReal* benchmark. We detail the standardized inference setup and the automated evaluation pipeline, emphasizing how specific prompting strategies were employed to elicit deep reasoning and quantify scientific accuracy.

7.1. Experimental Setup

To evaluate the depth of scientific reasoning, we moved beyond standard question-answering setups by enforcing a strict Chain-of-Thought (CoT) protocol. As illustrated in **Figure 7**, we designed a unified system prompt that mandates a dual-stage output format.

Enforcing Explicit Reasoning. The prompt explicitly instructs all MLLMs to separate their internal thought process (enclosed in `<think>` tags) from their final conclusion (enclosed in `<answer>` tags). This structural constraint serves two purposes:

1. **Qualitative Diagnosis:** It allows us to inspect the model's intermediate logic, identifying whether an error stems from a hallucinated visual feature or a flaw in reasoning.
2. **Persona Alignment:** By defining the role of a "polymer science expert," the prompt conditions the model to adopt a professional tone and prioritize scientific rigor over conversational casualness.

7.2. Automated Evaluation Pipeline

Given the open-ended nature of the tasks, we implemented a fine-grained "LLM-as-a-Judge" framework. Unlike simple keyword matching, our pipeline decouples the assessment of Completeness (Recall) from Correctness (Precision) using two distinct evaluator prompts.

Recall and Quality Assessment. As detailed in **Fig-**

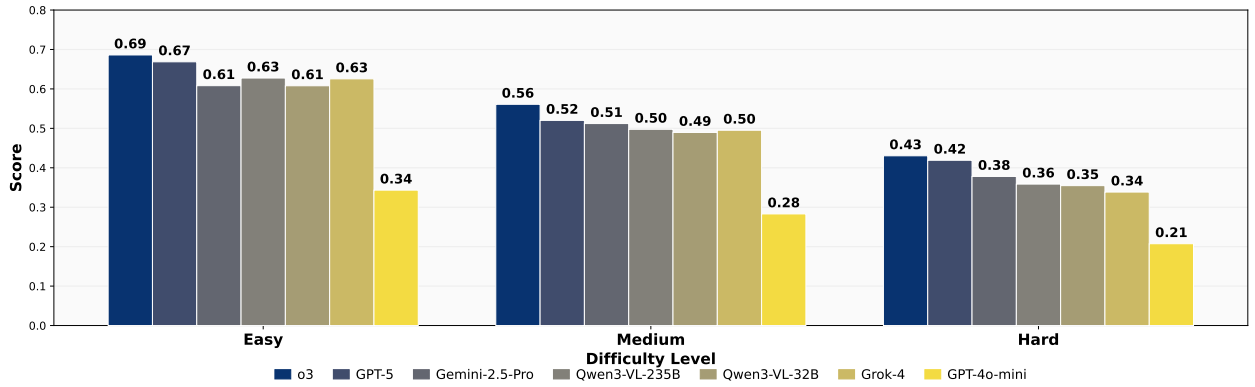

Figure 6. **Performance comparison across difficulty levels.** F1-scores for representative models are reported across Easy, Medium, and Hard subsets. A consistent performance degradation is observed as task complexity increases, with the "Hard" subset effectively differentiating robust state-of-the-art models (e.g., o3) from lightweight counterparts (e.g., GPT-4o-mini) which experience steeper declines.

Prompt for PolyReal Inference

Role: You are a polymer science expert. Your task is to provide a clear and accurate answer.

Process:

- **Internal Reasoning** (inside the `<think>` tag): Lay out your step-by-step reasoning process here.
- **Final Synthesized Answer** (inside the `<answer>` tag): After your reasoning, place the well-organized, clear, accurate, and concise answer within the tag. This answer must be a standalone, concise, and professional explanation that directly addresses the user's question. Do not simply repeat the reasoning process. You should distill the key conclusions from your thinking process to form a polished response.

Constraint: Please ensure your final response includes both the complete `<think>` and `<answer>` sections.

Figure 7. The standardized system prompt used for zero-shot inference across all MLLMs, enforcing a structured Chain-of-Thought reasoning process.

Automated Evaluator Prompt: Recall

Role: You are an exceptionally strict, meticulous, and critical grader specializing in polymer science. Your task is to evaluate a "Model's Answer" based on a list of "Key Scoring Points" and a "Ground Truth Answer".

Evaluation Task: You must evaluate *each* "Key Scoring Point" *sequentially* based on the following two dimensions:

1. Completeness (met - Binary Score):

- Does the "Model's Answer" clearly and unambiguously cover the core concept of this "Key Scoring Point"? This is a strict binary (0 or 1) check.
- **1 (Met):** The point is clearly and directly addressed, and its explanation has no critical information missing compared to the relevant explanation of this point in the "Ground Truth Answer".
- **0 (Not Met):** The point is missing, glossed over, vaguely implied, or has significant information omissions compared to the "Ground Truth Answer".

2. Professional Quality (quality_score - Float 0.0 to 1.0):

- **If met is 0**, this score **MUST** be **0.0**.
- **If met is 1**, you must then grade *how well* the point was covered:
  - **1.0 (Perfect):** The explanation is impeccable in fact, depth, and accuracy. The terminology is professional, the logic is rigorous.
  - **0.5 (Average):** The point is covered, but the explanation is superficial, imprecise, logically flawed, uses casual language, or is vague.
  - **0.1 (Poor):** The point is mentioned, but its explanation contains severe factual errors or logical fallacies.

Your Evaluation (in JSON format only): Output a SINGLE JSON object containing exactly three keys:

- **"met"**: [List of integers 0 or 1]
- **"quality_score"**: [List of floats between 0.0 and 1.0]
- **"reasoning"**: "A step-by-step, critical explanation. If 'Not Met', clearly specify the failure."

Figure 8. The full prompt template for the Precision Evaluator. We explicitly categorize distinct types of noise (highlighted in red) to rigorously penalize hallucinations and ensuring the Precision metric reflects true information density.

ure 8, the Recall Evaluator acts as a "strict grader." It verifies whether the model's response covers the expert-defined Key Scoring Points. Crucially, this prompt goes beyond a binary check; it includes a `quality_score` (0.0–1.0) to penalize correct but superficial explanations. This ensures that models are rewarded not just for mentioning a keyword, but for demonstrating a deep understanding of the underlying concept.

Precision and Noise Filtering. To address the "defensive verbosity" issue (where models generate excessive text to mask uncertainty), we designed a specialized Precision Evaluator shown in **Figure 9**. This prompt explicitly categorizes "False Positive" (FP) information into four distinct types: *Irrelevant*, *Incorrect*, *Redundant*, and *Filler*.

Automated Evaluator Prompt: Precision

Role: You are a rigorous, fair, and professional Benchmark Evaluator. Your task is to calculate the "Precision" of a "Model's Answer" and verify its coverage of the "Key Scoring Points."

Core Calculation Formula: Precision $= \frac{\text{Total TP Count}}{\text{Total TP Count} + \text{Total FP Count}}$

Evaluation Criteria: You must strictly adhere to the following definitions to count TP and FP:

- **TP (True Positive):** A specific unit of information (a phrase or sentence) within the "Model's Answer" that clearly and directly corresponds to one of the points in the "Key Scoring Points (Keywords)" list.
- **FP (False Positive):** A specific unit of information within the "Model's Answer" that falls into any of the following categories:
  - **[FP-Irrelevant]:** The information is correct but irrelevant to the current question.
  - **[FP-Incorrect]:** A factual error or hallucination.
  - **[FP-Redundant]:** A verbose, repetitive restatement of the same point already counted as a TP.
  - **[FP-Filler]:** "Empty" phrases with no informational value (e.g., "This is a good question," "In conclusion").

Note: Points from the Key Scoring Points list that are missed (False Negatives) do not participate in the Precision calculation.

Your Evaluation (in JSON format only): Your ONLY task is to identify all TP and FP information units. Output a single JSON object:

- **"tp_string"**: "(Found first TP unit)..."
- **"fp_string"**: "(Found first FP unit - [FP-Type])..."

Figure 9. Full prompt for the Precision Evaluator. The explicit categorization of False Positives (highlighted in red) ensures that verbose or hallucinatory content is accurately penalized.

By strictly penalizing these noise categories, we obtain a precise measure of the signal-to-noise ratio, distinguishing models that truly "know" the answer from those that simply "guess and expand."

7.3. Human Annotation Guidelines

Complementing the data sourcing strategy detailed in Section 3.1 of the main paper, this section outlines the rigorous annotation protocols followed by our expert team. To ensure the benchmark's reliability, we established a standardized workflow for converting raw scientific data into structured evaluable tasks.

Expert-Driven Curation. The annotation process was conducted exclusively by domain experts with deep specialization in polymer science, encompassing sub-fields such as macromolecular chemistry, spectroscopy, and materials engineering. Unlike general-purpose benchmarks, no non-expert annotators were involved. This ensured that the interpretation of complex instrument data (e.g., assigning NMR signals or analyzing reaction mechanisms) adhered to professional scientific standards.

Guideline for "Key Points" Extraction. To support the fine-grained evaluation metrics described in Section 7.2, experts were required to structure the Ground Truth Answer not as a monolithic block of text, but as a sequence of verifiable facts.

- **Atomic Decomposition:** Annotators decomposed complex reasoning chains into independent "Key Scoring Points." For example, in a spectral analysis task, identifying a specific peak wavenumber constitutes one point, while correctly assigning it to a functional group constitutes another.
- **Fact Verification:** Each key point was required to be explicitly supported by visual evidence in the provided data or by established chemical principles, strictly prohibiting ambiguous or subjective statements.

Cross-Validation and Quality Assurance. We implemented a "Peer-Review" validation protocol mirroring academic publication standards. After the initial authoring, each sample underwent a blind review by a second expert who attempted to solve the problem solely based on the visual input. Samples were flagged for revision if the second expert identified factual inaccuracies, ambiguous visual cues, or if the "Key Scoring Points" were insufficient to cover a complete scientific answer. Only samples achieving consensus were retained in the final dataset.

8. Additional Qualitative Analysis

A recurring failure mode observed in *PolyReal* is "Scientific Hallucination," where models generate plausible-sounding but factually non-existent evidence. Unlike general-domain hallucinations, these errors in polymer science typically stem from a conflict between the model's internal chemical priors and the specific visual data provided. We categorize these into two primary types:

8.1. Prior-Dominated Hallucination

This type occurs when the model prioritizes textual hints over visual evidence, leading to a "force-fit" analysis.

- **Mechanism:** When provided with a strong textual cue like a molecular formula (e.g., "C_2H_6OS" in Figure 4, Case 2), the model acts on statistical intuition rather than visual perception. It retrieves a plausible isomer from its internal knowledge base and attempts to rationalize the visual spectrum to match this pre-conceived structure.
- **Manifestation:** In the referenced NMR analysis case, the model hallucinated the existence of specific spectral peaks to justify its structural prediction. Despite the visual data showing a simpler signal pattern, the model fabricated descriptions of non-existent signals to maintain consistency with its initial hypothesis. This demonstrates a critical weakness in cross-modal alignment, where the model effectively "sees what it expects to see" rather than what is actually present.

8.2. Logical Over-Inference

This type arises when models bridge visual gaps using semantic probability, particularly in high-stakes scenarios like Lab Safety.

Lab Safety Analysis

Question:
You are operating on a 'polymerization completion processing' experimental bench... This bench contains large quantities of solvents commonly used in polymer synthesis... containing severe safety hazards... Please identify and analyze their causes.

Answer:
Large-capacity (several liters) 'ethyl acetate' (highly flammable) and 'N,N-dimethylformamide' (DMF, flammable and possessing reproductive toxicity/hepatotoxicity) are directly piled on the workbench... ethyl acetate vapors are extremely susceptible to ignition... meanwhile, the toxic vapors of DMF will continuously volatilize... Potential risk of mixed storage of incompatible chemicals... there is an extremely high risk of incompatible chemicals (such as strong oxidizers and flammable solvents) being mistakenly placed together, which could lead to exothermic reactions, fires, or the release of toxic gases.

Key Points:
- Accurately identify large-capacity flammable solvent 'ethyl acetate' on the bench.
- Accurately identify large-capacity toxic solvent 'N,N-dimethylformamide (DMF)'.
- Clearly point out the core mistake: large-capacity solvents should be stored in safety cabinets, not on the bench.
- Analyze the severe fire and explosion risk of flammable solvent (ethyl acetate) exposed.
- Analyze the acute or chronic poisoning health risk of toxic solvent (DMF) exposed.
- ....

Figure 10. Qualitative analysis of a failure case. Specific chemical substances are highlighted in blue, and identified safety hazards are highlighted in red.

Case Analysis To concretely illustrate this, we analyze an excerpt from the SOTA model O3's response to the Safety task (refer to the visual input in Figure 10). While O3 correctly identified the major hazards, its supporting evidence contained significant fabrications driven by a "Safety Alignment Bias"

O3 Response:

> *"Major catastrophic-level safety hazards... Several 5 L jerrycans of ethyl acetate are kept open or loosely capped on the open bench... Ignition sources are present (motor on the pressure reactor, possible heating mantle, metal tools, static discharge from plastic)... If over-pressurised, mechanically damaged or its relief valve vents, it can violently eject hot monomer/solvent aerosol."*

This excerpt reveals how the model "over-interprets" visual cues to align with a cautious safety narrative:

- **Visual Distortion (Blue):** The model correctly identifies the object ("Ethyl Acetate") but distorts its attributes to support a "Worst-Case Scenario." It exaggerates quantity ($1 \rightarrow$ *Several*) and misinterprets the state ($Closed \rightarrow$ *Kept Open*). By hallucinating that the containers are open, the model constructs a more compelling justification for the "Fire Risk," prioritizing a conservative safety warning over factual visual reporting.
- **Associative Hallucination (Red):** The model invents objects like a "*heating mantle*" or "*relief valve*" which are visually absent. These hallucinations stem from **Probabilistic Semantic Association**: since "Reactors" co-occur with "Heating Mantles" in chemical safety literature, the model infers their presence to complete a logical "Ignition Source" chain.

Impact. This pattern poses a subtle risk: **Correct Conclusion, Fabricated Evidence.** While the safety warning is valid, the user might be misled to search for non-existent hazards (e.g., a heating mantle) while overlooking real ones. This necessitates the rigorous *Key Scoring Point* verification protocol used in *PolyReal* to penalize invented data.

9. Small-Model Results

To complement the main results, we additionally evaluate smaller models in the 2B–13B range. As shown in Table 7, the results exhibit a clear scaling trend: performance improves substantially with model size. Moreover, at comparable scales, reasoning-oriented models generally outperform their standard counterparts, suggesting that explicit reasoning remains beneficial in the small-model regime.

Table 7. Performance of additional small models on PolyReal.

Model	Precision (Avg.)	Recall (Avg.)	F1 (Avg.)
LLaVA-v1.6-Vicuna-13B	0.110	0.058	0.053
Qwen3-VL-2B-Thinking	0.213	0.196	0.166
InternVL3_5-8B	0.415	0.366	0.337
Qwen3-VL-8B-Instruct	0.543	0.345	0.366
Qwen3-VL-8B-Thinking	0.442	0.442	0.395