# DisastQA: A Comprehensive Benchmark for Evaluating Question Answering in Disaster Management

**Anonymous ACL submission**


## Abstract

Accurate question answering (QA) in **disaster management** requires reasoning over uncertain and conflicting information, a setting poorly captured by existing benchmarks built on clean evidence. We introduce DISASTQA, a large-scale benchmark of 3,000 rigorously verified questions (2,000 multiple-choice and 1,000 open-ended) spanning eight disaster types. The benchmark is constructed via a human–LLM collaboration pipeline with stratified sampling to ensure balanced coverage. Models are evaluated under varying evidence conditions, from closed-book to noisy evidence integration, enabling separation of internal knowledge from reasoning under imperfect information. For open-ended QA, we propose a human-verified keypoint-based evaluation protocol emphasizing factual completeness over verbosity. Experiments with 20 models reveal substantial divergences from general-purpose leaderboards such as MMLU-Pro. While recent open-weight models approach proprietary systems in clean settings, performance degrades sharply under realistic noise, exposing critical reliability gaps for disaster response. All code, data, and evaluation resources are available at the project page.


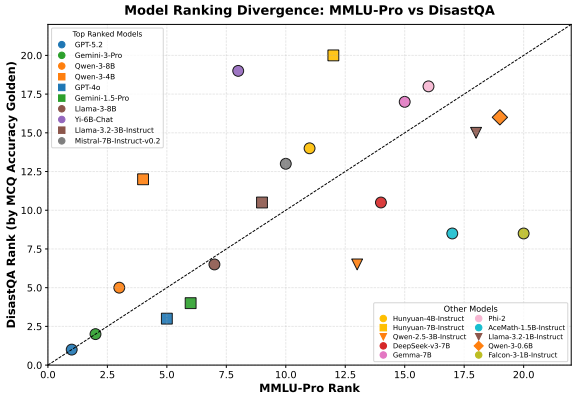


Figure 1: Model ranking divergence between DisastQA-MCQ (Gold) and the MMLU-Pro subset. Pronounced off-diagonal deviations indicate that general QA leaderboards may not reliably predict relative performance in high-stakes disaster-response QA. Full numerical results are provided in Table 13.

## 1 Introduction

Disasters inevitably trigger severe societal, economic, and humanitarian consequences. Reliable question answering (QA) systems crucially support timely decision-making, situational awareness, and coordinated responses during crises (Palen and Anderson, 2016; Vieweg et al., 2010). Large Language Models (LLMs) hold great promise for advancing such capabilities by leveraging both their *parametric knowledge* and *external evidence*. However, with the rapid evolution of LLMs, a key question arises: *can even the latest frontier models maintain reliability in these high-stakes scenarios?*

Disaster-response QA differs fundamentally from general-domain QA. Information in this domain is highly fragmented across heterogeneous sources, ranging from official bulletins and scientific reports to unstructured social media streams (Imran et al., 2015; Alam et al., 2021). These sources are often incomplete, noisy, or contradictory, and answering disaster-related questions requires synthesizing multiple factual aspects such as hazard type, affected regions, and resource availability. This makes *factual completeness and accurate information integration* essential for reliability.

However, existing QA and evaluation benchmarks mainly target general knowledge (Rajpurkar et al., 2016; Joshi et al., 2017; Hendrycks et al., 2021) or domains such as medicine and law (Jin et al., 2019; Zhong et al., 2024), leaving disaster contexts underexplored. Furthermore, most existing datasets are *multiple-choice (MCQ)* based, failing to assess reasoning depth and factual completeness. In disaster management, decision-makers often require *open-ended, evidence-grounded responses*, e.g., describing affected regions, resource needs, or response measures, rather than selecting

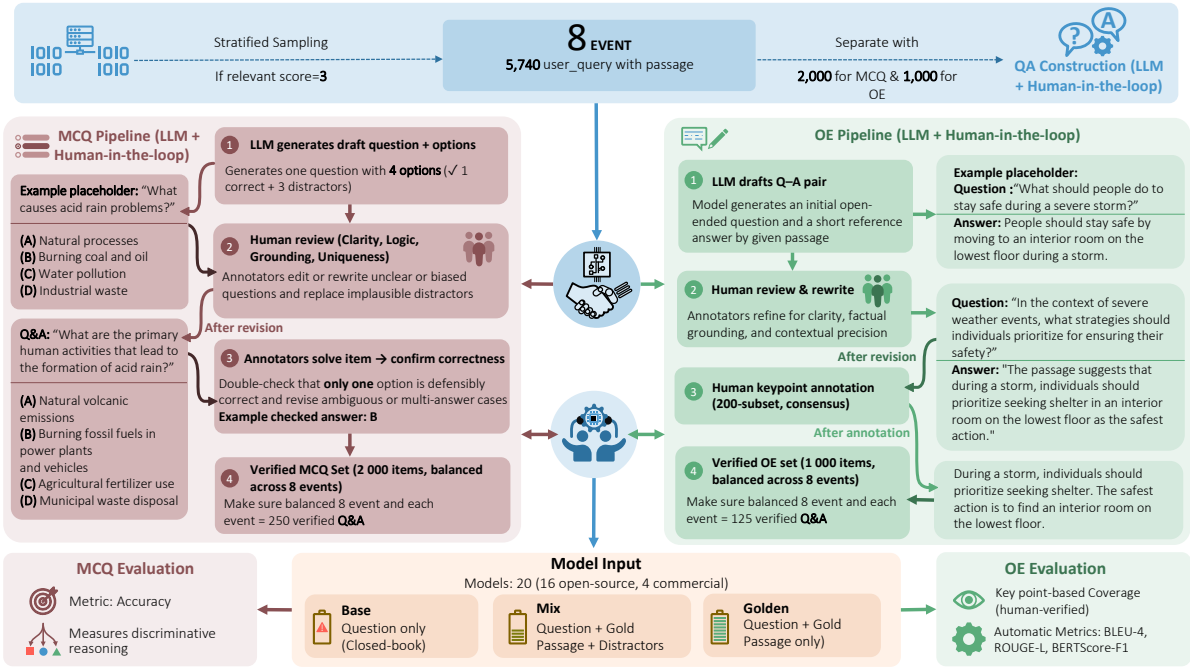


Figure 2: The **Human–LLM Collaborative Pipeline** for DisastQA. User queries are stratified across 8 disaster event types (**Top**). Two parallel tracks construct the dataset: the **MCQ Pipeline** with human refinement for distractor quality and answer correctness, and the **OE Pipeline** with human rewriting and **Keypoint** annotation (**Middle**). Models are evaluated under three evidence settings (*Base*, *Mix*, *Golden*) (**Bottom**); see Section 5 for details.

from fixed options. Thus, existing QA benchmarks fall short in evaluating reliability and completeness under decision-critical conditions.

To address these gaps, we introduce DisastQA, the first large-scale benchmark for disaster-response QA integrating both Multiple-Choice (MCQ) and Open-Ended (OE) tasks. Developing such a benchmark poses two main challenges: (1) evaluating open-ended responses in an interpretable and reliable way that captures factual completeness, and (2) constructing large-scale, high-quality QA data without prohibitive manual annotation costs.

To tackle these challenges, DisastQA adopts a **Human–LLM collaboration pipeline** (Figure 2) that balances scalability with rigorous human oversight, ensuring high-quality, factually grounded, and diverse QA pairs across eight major disaster categories. We explicitly evaluate models under distinct information settings—ranging from closed-book generation to reasoning with noisy retrieval results. This design disentangles a model's internal knowledge from its ability to reason with external information. For open-ended QA, we introduce **Keypoint Coverage**, a metric designed to measure factual completeness beyond lexical overlap, addressing the limitations of surface-level metrics such as ROUGE and BLEU. Together, these designs make DisastQA a systematic framework for evaluating both knowledge-based and evidence-grounded capabilities of LLMs.

Using DISASTQA, we evaluate 20 open-source and commercial LLMs, including frontier systems such as GPT-5.2 and Gemini-3 Pro. We compare against a general-domain benchmark (MMLU-Pro) and assess factual coverage in OE tasks. Our results reveal two critical insights: (1) Closing Performance Gaps: Efficient open-weight models such as Qwen-3-8B now approach proprietary leaders (e.g., GPT-4o), indicating a narrowing capability gap in this domain. (2) The Persistence of Reasoning Gaps: Even the strongest frontier models degrade under retrieval noise (i.e., top-$k$ retrieved contexts containing both relevant evidence and plausible distractor passages) and fail to achieve perfect Keypoint Coverage in complex scenarios. These findings highlight that while scaling improves general capabilities, domain-specific reliability in safety-critical contexts remains an unsolved challenge.

Our Contributions. We (1) construct the first large-scale, human-verified benchmark for disaster QA through a Human-LLM collaboration pipeline; (2) introduce a transferable keypoint-based evaluation framework to measure factual completeness in open-ended QA; and (3) conduct comprehensive

evaluations of 20 LLMs, establishing new upper bounds with frontier models while revealing persistent reliability gaps in disaster-specific reasoning.

## 2 Related Work

**Existing QA Benchmarks.** Standard QA benchmarks emphasize broad coverage but **overlook** the reliability constraints of disaster-response settings. For example, SQuAD-style reading-comprehension datasets (Rajpurkar et al., 2016) **assume relatively clean**, single-context evidence, missing the uncertainty common during crises. Large-scale long-form QA datasets such as ELI5 (Fan et al., 2019) and GOOAQ (Khashabi et al., 2021) rely on web documents, while WikiCQA (Dong et al., 2023) targets closed-book settings. Meanwhile, benchmarks such as MMLU-Pro (Wang et al., 2024) mainly probe parametric knowledge through multiple-choice questions, **limiting evaluation of** evidence-grounded, open-ended synthesis. As a result, neither paradigm directly tests robustness to retrieval noise or the *factual completeness* required for decision support.

**Disaster-Focused Resources.** Prior work in disaster informatics has focused on information extraction tasks such as classification, retrieval, and summarization (Imran et al., 2015; Olteanu et al., 2014; Rudra et al., 2019), rather than generative QA. While a few attempts like DisasterQA (Rawat, 2024) exist, they are constrained by small scale, reliance on multiple-choice formats, and a lack of controlled uncertainty evaluation. **DisastQA** bridges these gaps by integrating both Multiple-Choice and Open-Ended tasks under a tri-level evaluation setup (*Base*, *Mix*, *Golden*) to disentangle internal knowledge from evidence reasoning. Furthermore, we introduce a *keypoint-based protocol* to rigorously quantify factual completeness, offering the first large-scale, reproducible, and evidence-grounded benchmark for the domain.

## 3 DisastQA: Disaster Management Question Answer Benchmark

### 3.1 Overview

We introduce **DisastQA**, a large-scale benchmark for disaster-response QA comprising two complementary tasks: DisastQA-MCQ for multiple-choice and DisastQA-OE for open-ended questions. These tasks collectively span eight diverse real-world disaster event types with balanced coverage across categories (details in Appendix A).

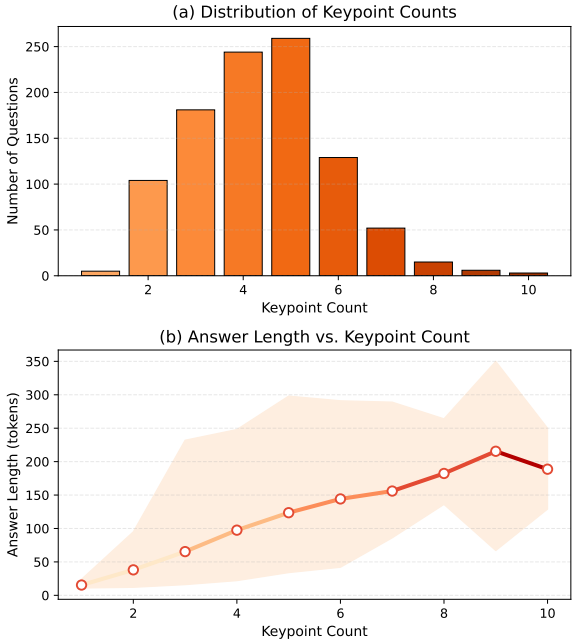


Figure 3: (a) Distribution of keypoint counts per OE answer. (b) Relationship between answer length (in tokens) and keypoint count, showing that longer answers typically contain more keypoints.

MCQ items evaluate discriminative reasoning by requiring models to identify the correct answer among plausible, human-authored distractors, whereas OE items assess factual completeness through free-form responses reflecting real-world information needs in disaster management.

Construction follows a Human–LLM Collaboration Pipeline (Figure 2) and is formally summarized in Algorithm 1. This hybrid approach strategically balances scalability with the factual rigor required for a safety-critical domain, consistent with prior visions of collective human–machine intelligence (Li et al., 2025). While LLMs provide scalable initial drafts, rigorous human expert oversight remains indispensable for rewriting, verification, and ensuring factual trustworthiness (see Appendix B for detailed error analysis and refinement statistics). This design, detailed in the following sections, ensures DISASTQA is both large-scale and reliable.

### 3.2 Annotator Expertise

All annotation and verification were performed by a team of five researchers with backgrounds in disaster resilience, crisis informatics, and infrastructure systems. The annotators possessed deep expertise across multiple disaster domains, ensuring accurate interpretation of diverse event types. Such

domain-specific knowledge is essential for both verifying the factual accuracy of technical content and authoring plausible, context-aware distractors. They oversaw the entire quality-control process as well as the final validation of all items. Detailed annotation procedures for MCQ and OE tasks are described in Sections 3.3 and 3.4.

## 3.3 DisastQA-MCQ Construction

DisastQA-MCQ is constructed through a three-stage Human–LLM collaboration framework, employing an **evidence-grounded generation strategy** with strict factual constraints to convert disaster-related search queries into high-quality multiple-choice questions. We leverage the corpus from DisastIR (Yin et al., 2025), a domain-specific benchmark for crisis information seeking, primarily as a source of authoritative evidence passages, ensuring that all questions are rooted in real-world disaster reports rather than synthesized abstractions.

### 3.3.1 Query-conditioned Question Generation

Since the raw source queries are often short and keyword-based (e.g., "earthquake bridge collapse", "flood evacuation route"), they are unsuitable for direct use in QA. To address this, we prompt an LLM to rewrite each query into a well-formed, information-seeking question, leveraging synthetic generation techniques for domain adaptation (Ma et al., 2021). These rewritten questions are conditioned on passages annotated as highly relevant (i.e., a score of 3 on a 0–3 scale). Selecting only these top-tier passages ensures that rewritten questions are structurally grounded in definitive evidence rather than tangential mentions. This step guarantees balanced coverage across the eight disaster categories and establishes a reliable factual basis for subsequent answer generation.

### 3.3.2 Answer Option Generation

For each rewritten question–passage pair, the LLM generates one correct answer directly supported by the passage and three distractors that are semantically plausible but factually incorrect. These distractors are initially produced via strategic alteration of critical factual attributes (e.g., event date, affected location, magnitude, or policy detail). This stage captures the model's ability to generate diverse and competitive options, forming the preliminary multiple-choice structure. However, our pilot analysis revealed that LLM-generated distractors, while scalable, often lack sufficient discriminative power or remain trivially "easy" (e.g., simple negations, contextually irrelevant options). This finding underscores the necessity of the subsequent human refinement stage (Section 3.3.3) to systematically re-author and verify these options, ensuring all distractors are both challenging and realistic.

### 3.3.3 Human Expert Refinement

Human experts then refine all question–answer sets to ensure factual accuracy, clarity, and difficulty balance. Annotators verify that the gold answer is strictly supported by the passage and **manually revise distractors to ensure they are semantically plausible yet factually incorrect** (see Table 9 in Appendix D.1 for examples). Each MCQ is independently validated by two annotators to confirm that exactly one option is correct; any disagreement is resolved by consensus. Additionally, correct answers are uniformly distributed across the four option slots (A/B/C/D) to eliminate positional bias.

This rigorous refinement ensures that DisastQA-MCQ evaluates factual discrimination rather than mere surface-level memorization. Quality control analysis reveals a human verification consistency of 95.4% (954/1000 sampled items), where human experts independently solved the questions and matched the generated gold answers. This high consistency confirms that the distractors are unambiguous and the gold answers are objectively correct. Detailed prompt templates, filtering statistics, and annotation guidelines are provided in Appendix E.2.2 and E.1.

## 3.4 DisastQA-OE Construction

### 3.4.1 Query-conditioned Question Generation

Similar to the MCQ track, the construction of OE questions begins with a DisastIR query–passage pair $(q, p_g)$, where $p_g$ is the most relevant passage (relevance score = 3). The LLM first rewrites the query into a natural-language question grounded in $p_g$. Human annotators then rigorously refine this draft for clarity, answerability, and factual fidelity. Unlike the MCQ track, this stage omits option generation and focuses solely on producing a concise, unambiguous question that elicits an explanatory, evidence-grounded response. Prompt templates for this stage are listed in Appendix I.

### 3.4.2 Answer Construction

For each verified question, annotators composed a gold reference answer ($a_{ref}$) grounded in the

| Statistic | MCQ | OE |
|---|---|---|
| # Instances | 2,000 | 1,000 |
| Question Length | 15.1±3.3 [5,32] | 31.1±10.4 [7,78] |
| Option Length | 9.5±3.3 [2,26] | – |
| Answer Length | – | 103.4±52.5 [11,350] |
| Keypoint Count | – | 4.4±1.5 [1,14] |

Table 1: Descriptive statistics for the DisastQA-MCQ and DisastQA-OE tasks. All lengths are reported in tokens as mean±std [min,max].

passage. While LLMs were optionally used to generate preliminary drafts, all reference answers were rewritten and verified by human experts to ensure high factual accuracy, conciseness, and neutrality. This approach maintains pipeline consistency with the MCQ track while also eliminating residual model bias or any unsupported content, yielding reliable and verifiable gold references, which in turn serve as the foundation for subsequent keypoint decomposition and fact-level evaluation.

#### 3.4.3 Keypoint Annotation

To enable fine-grained factual evaluation, a 200-item subset of reference answers was decomposed into minimal factual units called **keypoints** ($\mathcal{K}$). For instance, a reference answer describing emergency safety actions is decomposed into atomic units such as "(1) issue interim safety guidance" and "(2) coordinate with local fire services," as illustrated in Table 15 (Appendix). Since these keypoints serve as the ground truth for our evaluation metric, we employed a fully manual annotation process to ensure maximum reliability and eliminate LLM-induced biases or inconsistencies.

To ensure reliability without relying on single-annotator subjectivity, we adopted a **strict double-pass protocol with consensus resolution**. Specifically, all reference answers were first annotated by primary experts and then reviewed by a verifier. Any discrepancies in keypoint boundaries were resolved through discussion to establish the final ground truth. Statistical analysis of the final dataset shows a consistent granularity with an average of 4.39 keypoints per answer (SD=1.55), confirming stable annotation standards across disaster types (see Appendix E.2.1 for detailed annotation guidelines). This annotation infrastructure underpins our primary metric, **Keypoint Coverage** (formally defined in Section 5.3), which evaluates models based on their recall of these essential atomic facts rather than surface-level lexical overlap.

## 4 Dataset Statistics and Analysis

We analyze the statistical properties of DisastQA, focusing on the impact of human refinement, descriptive trends, and domain complexity.

### 4.1 Descriptive Statistics

The final DISASTQA dataset comprises 3,000 instances, including 2,000 multiple-choice and 1,000 open-ended questions, with balanced coverage across eight disaster types (see Table 1). MCQ items are precision-oriented (avg. 15.1 tokens), assessing discriminative reasoning over specific facts. In contrast, OE items feature longer questions (avg. 31.1 tokens) that require comprehensive, human-authored reference answers (avg. 103.4 tokens). Within the annotated subset, each reference answer is decomposed into an average of 4.4 atomic keypoints, serving as a quantifiable measure of factual complexity and granular information density.

### 4.2 Domain-Specific Complexity Analysis

The keypoint annotations quantitatively reveal the high factual density that characterizes disaster-response QA. As shown in Figure 3, while many questions require reasoning over 3–5 facts, a significant portion necessitate integrating 8 or more atomic facts, often dispersed throughout the context, representing a level of complexity not typically captured in general-domain QA benchmarks.

This high factual density underscores the distinct challenges of the disaster domain. This validates the necessity of our keypoint-based evaluation protocol, as traditional metrics such as ROUGE often fail to penalize models that produce superficially fluent yet factually incomplete answers, a limitation widely recognized in recent generation benchmarks (Scialom et al., 2021).

## 5 Evaluation Methodology

We denote $a^\star$ as the gold answer for MCQ, $\mathcal{K}$ as the gold keypoints for OE, and $a_{gen}$ as the model-generated answer. To ensure systematic and interpretable evaluation, we adopt a unified methodology disentangling a model's parametric knowledge from its ability to utilize external evidence. Our core innovation is a three-context evaluation framework comprising *Base*, *Mix*, and *Golden*, which builds upon prior work on robustQA (Han et al., 2023). This framework allows us to assess model behavior not only under idealized conditions but also under realistic uncertainty, reflect-

ing how disaster information is often incomplete, noisy, or misleading. We compare these results with general-domain benchmarks such as Natural Questions (Kwiatkowski et al., 2019) (Section 6) to contextualize disaster-domain robustness.

### 5.1 Evaluation Settings

To disentangle parametric knowledge from reasoning capabilities, we evaluate models under three controlled information settings (see qualitative examples in Appendix I):

**Base (Closed-Book).** The model answers the question $q$ using only its internal parametric knowledge, serving as a baseline for hallucination and internal knowledge reliability.

**Golden (Oracle).** The model receives the question $q$ and ground-truth passage $p_g$, representing an oracle condition with perfect evidence and measuring upper-bound reasoning performance.

**Mix (Noisy Context).** The context $p_c$ combines the golden passage $p_g$ with $k = 4$ stratified distractors ($P_{dist}$), simulating a standard **Top-5 retrieval** scenario (1 Gold + 4 Noise) (Karpukhin et al., 2020; Chen et al., 2024). Rather than evaluating retrieval performance, we quantify the LLM's discriminative ability to identify and utilize relevant evidence under high-ranking noise. Passages are presented in randomized order to prevent position bias.

### 5.2 MCQ Evaluation

Each MCQ instance consists of a question $x$, four options $\{o_A, o_B, o_C, o_D\}$ with one correct answer $a^\star$, and a passage $p_c$ under the chosen evaluation context. Models are required to output the option.

**Accuracy Metric.** Evaluation is based on *exact-match accuracy*, i.e., the proportion of test questions for which the predicted option matches the gold answer. Since each item is validated to ensure exactly one correct answer, accuracy directly reflects a model's ability to discriminate correctness.

### 5.3 OE Evaluation

Each OE instance consists of a question $x$, a passage $p_c$, and a human reference answer $a_{ref}$. To rigorously evaluate factual adequacy, we adopt a **Human-Verified Keypoint Protocol**.

**Keypoint Decomposition.** Domain experts decompose each reference answer $a_{ref}$ into a set of atomic keypoints $\mathcal{K} = \{k_1, k_2, \ldots, k_n\}$, aligning with the atomic fact definition in FActScore (Min et al., 2023), where each keypoint represents a minimal factual unit required for correctness.

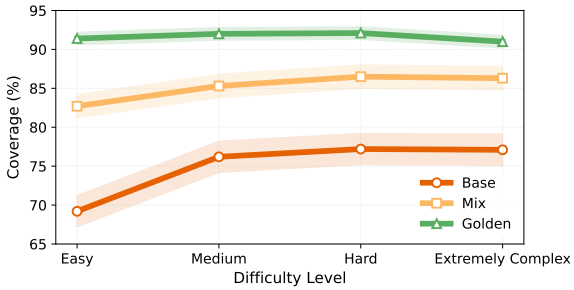


Figure 4: Keypoint Coverage across difficulty levels under Base, Mix, and Golden.

**Scoring Procedure (Human Verification).** We conduct a fully human-annotated evaluation on a representative 200-item OE subset. To ensure structural balance and representativeness, we employ **Stratified Random Sampling**: we randomly sample exactly 25 questions from each of the 8 disaster types (Total: $25 \times 8 = 200$). All 20 models are evaluated on this identical subset to guarantee fair, paired comparisons. Annotators are presented with the model-generated answer $a_{gen}$ and the gold keypoints $\mathcal{K}$. For each keypoint $k_i \in \mathcal{K}$, the annotator determines whether its semantic meaning is entailed by $a_{gen}$, regardless of surface phrasing. This binary judgment $\mathbb{I}(k_i \in a_{gen})$ equals 1 if the fact is present and 0 otherwise. Detailed human annotation guidelines are provided in Appendix E.2.2.

We define Keypoint Coverage as a metric of strict factual recall:

$$\text{Coverage}(a_{gen}) = \frac{1}{|\mathcal{K}|} \sum_{k \in \mathcal{K}} \mathbb{I}(k \in a_{gen}). \quad (1)$$

This human-verified evaluation prioritizes factual completeness and avoids rewarding verbosity or surface-level lexical overlap, making it particularly suitable for risk-sensitive disaster QA settings.

**Complementary Metrics.** For comparability with prior work, we report ROUGE-L, BLEU-4, and BERTScore-F1 on all 1,000 OE questions. These automated metrics are secondary indicators of surface-level similarity, while human-verified Keypoint Coverage serves as the primary measure of factual adequacy.

### 5.4 Summary of Experimental Setup

In summary, DisastQA integrates **Accuracy** for MCQ and **Keypoint Coverage** for OE. Together

| Model | Params | MCQ | | | OE | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Accuracy | | | ROUGE-L | | | BLEU-4 | | | BERTScore-F1 | | | Coverage (%) | | |
| | | Base | Mix | Golden | Base | Mix | Golden | Base | Mix | Golden | Base | Mix | Golden | Base | Mix | Golden |
| GPT-5.2 | – | **0.9314** | **0.9670** | **0.9965** | 0.178 | 0.294 | 0.286 | 0.026 | 0.074 | 0.063 | 0.854 | 0.886 | 0.882 | **87.1** | **93.5** | 94.6 |
| Gemini-3 Pro | – | 0.9219 | 0.9535 | 0.9670 | 0.210 | **0.426** | 0.437 | 0.043 | 0.196 | 0.193 | 0.868 | **0.913** | 0.919 | 84.7 | 91.5 | **96.5** |
| GPT-4o | – | 0.9105 | 0.9625 | 0.9935 | 0.229 | 0.381 | 0.367 | 0.069 | 0.174 | 0.160 | 0.878 | 0.910 | 0.910 | 83.7 | 89.8 | 95.4 |
| Gemini-1.5 Pro | – | 0.8895 | 0.9380 | 0.9870 | 0.205 | 0.399 | 0.317 | 0.038 | 0.178 | 0.094 | 0.868 | 0.906 | 0.897 | 84.5 | 84.0 | 95.1 |
| Qwen-3-8B | 8.19B | 0.8865 | 0.9625 | **0.9965** | 0.234 | 0.332 | 0.357 | 0.063 | 0.129 | 0.139 | 0.878 | 0.896 | 0.906 | 79.1 | 87.1 | 94.0 |
| Yi-6B-Chat | 6.06B | 0.8575 | 0.9565 | 0.9930 | 0.257 | 0.305 | 0.358 | 0.090 | 0.126 | 0.167 | 0.879 | 0.892 | 0.900 | 73.0 | 84.7 | 86.5 |
| Llama-3-8B | 8.03B | 0.8715 | 0.9565 | 0.9910 | **0.273** | 0.378 | 0.425 | **0.104** | **0.202** | 0.223 | **0.885** | 0.902 | 0.917 | 77.5 | 85.0 | 93.4 |
| Qwen-3-4B | 4.02B | 0.8605 | 0.9600 | 0.9960 | 0.234 | 0.337 | 0.356 | 0.057 | 0.125 | 0.127 | 0.879 | 0.897 | 0.907 | 76.5 | 83.5 | 91.7 |
| Qwen-2.5-3B-Instr. | 3.00B | 0.8530 | 0.9545 | 0.9915 | 0.236 | 0.339 | 0.364 | 0.061 | 0.129 | 0.134 | 0.880 | 0.899 | 0.906 | 76.9 | 85.4 | 93.4 |
| Phi-2 | 2.78B | 0.8725 | 0.9570 | 0.9895 | 0.251 | 0.323 | 0.355 | 0.088 | 0.154 | 0.169 | 0.882 | 0.897 | 0.905 | 74.0 | 82.0 | 86.9 |
| Mistral-7B-Instr. | 7.24B | 0.8180 | 0.9435 | 0.9865 | 0.243 | 0.328 | 0.384 | 0.072 | 0.145 | 0.168 | 0.879 | 0.896 | 0.911 | 76.3 | 85.4 | 91.6 |
| Gemma-7B | 8.54B | 0.8025 | 0.8975 | 0.9805 | 0.263 | 0.386 | **0.467** | 0.088 | 0.189 | **0.242** | 0.879 | 0.907 | **0.922** | 72.9 | 69.7 | 89.7 |
| DeepSeek-v3-7B | 7.00B | 0.8020 | 0.9435 | 0.9860 | 0.270 | 0.377 | 0.422 | 0.100 | 0.193 | 0.221 | 0.884 | 0.906 | 0.915 | 73.5 | 84.1 | 92.0 |
| Llama-3.2-3B-Instr. | 3.21B | 0.8510 | 0.9550 | 0.9865 | **0.273** | 0.374 | 0.418 | 0.094 | 0.177 | 0.201 | 0.884 | 0.905 | 0.916 | 76.6 | 84.3 | 92.0 |
| Llama-3.2-1B-Instr. | 1.24B | 0.8120 | 0.9345 | 0.9690 | 0.256 | 0.326 | 0.376 | 0.091 | 0.156 | 0.190 | 0.881 | 0.897 | 0.909 | 74.5 | 86.5 | 90.4 |
| Falcon-3-1B-Instr. | 1.67B | 0.7475 | 0.9120 | 0.9690 | 0.248 | 0.343 | 0.401 | 0.078 | 0.157 | 0.174 | 0.882 | 0.900 | 0.914 | 71.1 | 84.1 | 92.3 |
| AceMath-1.5B-Instr. | 1.78B | 0.7410 | 0.9210 | 0.9815 | 0.244 | 0.342 | 0.373 | 0.082 | 0.167 | 0.179 | 0.873 | 0.899 | 0.897 | 67.9 | 86.6 | 92.3 |
| Hunyuan-4B-Instr. | 4.22B | 0.8195 | 0.9435 | 0.9780 | 0.192 | 0.278 | 0.284 | 0.043 | 0.106 | 0.100 | 0.861 | 0.885 | 0.884 | 67.2 | 84.5 | 91.4 |
| Hunyuan-7B-Instr. | 7.50B | 0.5700 | 0.8790 | 0.9305 | 0.123 | 0.323 | 0.266 | 0.012 | 0.127 | 0.083 | 0.816 | 0.892 | 0.881 | 39.6 | 72.6 | 79.9 |
| Qwen-3-0.6B | 0.75B | 0.7455 | 0.8290 | 0.8810 | 0.242 | 0.289 | 0.314 | 0.077 | 0.129 | 0.123 | 0.869 | 0.870 | 0.875 | 69.7 | 85.7 | 90.1 |

Table 2: Summary of model performance on DISASTQA-MCQ and OE tasks across three retrieval settings (Base, Mix, Golden). Metrics include Accuracy for MCQ, and ROUGE-L, BLEU-4, BERTScore-F1, and Keypoint Coverage for OE evaluation. Best results in each column are highlighted in **bold**.

with controlled retrieval settings, these metrics ensure that our benchmark captures (1) discriminative reasoning, (2) factual adequacy, and (3) evidence utilization under uncertainty. To ensure reproducibility, we adopt a deterministic decoding strategy for all generative tasks, eliminating randomness to enable fair comparison.

### 5.5 Evaluated Models

We evaluate 20 models, spanning diverse scales from efficient open-weight models (0.6B–8B) to proprietary frontier systems. We strictly prioritize instruction-tuned variants to align with conversational deployment needs. Our selection covers major open families (*Qwen, Llama, Gemma, Mistral, Phi, Falcon, Yi, DeepSeek, Hunyuan*) and proprietary APIs, including the latest state-of-the-art versions (*GPT-5.2*, *Gemini-3 Pro*). This selection strategy allows us to benchmark the ~7B parameter class, a standard for local resource-constrained disaster scenarios, against the performance upper bounds established by proprietary SOTAs, revealing the critical trade-off between practical deployment efficiency and robust reasoning reliability. Detailed implementation notes and model sources are provided in Appendix F.

## 6 Results and Analysis

### 6.1 Overall Performance on DisastQA-MCQ

Table 2 summarizes the performance of 20 models. Across all settings, we observe a consistent hierarchy of *Base* < *Mix* < *Golden*, confirming that DISASTQA questions require external evidence, underscoring the **critical role of context retrieval**.

**Gap between Open-Source and Proprietary Models.** In the *Base* setting, the latest frontier model GPT-5.2 establishes a new upper bound with 93.1% accuracy. Notably, the open-weight model Qwen-3-8B performs exceptionally well, narrowing the gap with proprietary systems in the *Golden* setting. However, in the *Base* (no-context) scenario, **GPT-5.2 retains a substantial advantage** (93.1% vs. Qwen-3's 88.7%), indicating that proprietary models still possess stronger internalized world knowledge derived from massive pre-training.

**Robustness to Contextual Noise.** Comparing *Golden* and *Mix* reveals how models handle noisy evidence. While GPT-5.2 achieves the highest absolute accuracy in *Mix*, **Gemini-3 Pro exhibits the strongest relative robustness**, showing the smallest drop from *Golden* to *Mix*. This indicates that although GPT-5.2 excels in precision, Gemini-3 Pro is particularly stable in filtering irrelevant or conflicting passages, an essential capability for real-world decision-support scenarios.

### 6.2 Overall Performance on DisastQA-OE

**Trade-off: Fluency vs. Factuality.** Results reveal a clear **trade-off between fluency and factual completeness**. As shown in Table 2, Gemma-7B achieves high ROUGE-L scores, reflecting text flu-

ency, yet its Keypoint Coverage lags behind. By contrast, Gemini-3 Pro demonstrates superior factual stability. While smaller models often struggle to maintain information density in longer responses, Gemini-3 Pro consistently synthesizes comprehensive, multi-fact answers. This finding is critical: whereas GPT-5.2 dominates discriminative tasks (MCQ), Gemini-3 Pro exhibits the generative capabilities required for high-fidelity disaster reporting.

### 6.3 Performance by Complexity

Figure 4 illustrates the aggregate performance trends across difficulty levels, while detailed per-model numerical results are provided in Table 20 (Appendix G). The dataset distribution is skewed (see Table 3), with "Extremely Complex" items representing a long-tail subset ($N = 26$). Due to this scarcity, results on this tier should be interpreted as exploratory trends rather than statistically conclusive benchmarks. However, the relative degradation trend offers a robust signal: While frontier models like GPT-5.2 and Gemini-3 Pro maintain high performance on complex queries, **smaller models degrade sharply**. For instance, Qwen-3-8B drops from 94.2% on Easy items to 87.5% on Extremely Complex ones in the *Golden* setting. This suggests that while smaller models are competent on simple facts, they lack the reasoning depth to integrate multiple evidence sources to derive correct conclusions. This fragility is exacerbated in the *Mix* setting, where **reasoning complexity amplifies the distraction effect**, causing a sharp performance degradation for sub-10B models.

### 6.4 Comparison with General-Domain Benchmarks

To test domain transfer, we compare DISASTQA-MCQ against an 8-subject subset of MMLU-Pro (Wang et al., 2024) (details in Appendix C). Figure 1 illustrates the ranking divergence. While GPT-5.2 leads on both, models like DeepSeek-v3-7B drop significantly in disaster contexts compared to general benchmarks. This confirms that **general-domain performance does not guarantee reliability** in safety-critical scenarios.

### 6.5 Event Type Breakdown

As illustrated in Appendix J (Figure 5), there is substantial performance variance across disaster types (Table 21). In the *Golden* setting, GPT-5.2 achieves near-saturated performance (>99%) across all categories. However, domain gaps remain for open-weight models. For instance, in the *Base* setting, DeepSeek-v3-7B drops to 78.8% accuracy on "Biological" vs. 82.0% on "Geohazards". **Technological and Biological disasters are consistently the most challenging** due to specialized terminology (e.g., chemical compounds, pathogen details) that is underrepresented in pre-training data.

| Difficulty | #Q | Proportion |
|---|---|---|
| Easy | 109 | 10.9% |
| Medium | 684 | 68.4% |
| Hard | 181 | 18.1% |
| Extremely Complex | 26 | 2.6% |
| Total | 1,000 | 100% |

Table 3: **Distribution of difficulty levels for open-ended questions.** Tiers are defined by keypoint (KP) counts: Easy (1–2 KP), Medium (3–5 KP), Hard (6–7 KP), and Extremely Complex (8+ KP).

### 6.6 Error Analysis

We conducted a qualitative analysis of failure cases (see Appendix B). For MCQs, the dominant failure mode is distractor attraction, where models are misled by incorrect options sharing high lexical overlap with the question. This reveals reliance on **spurious keyword matching rather than multi-hop verification**. For OEs, even GPT-5.2 occasionally exhibits partial coverage, omitting quantitative details (e.g., casualty figures) while capturing the general narrative. This underscores the value of Keypoint Coverage as an evaluation signal for identifying **factual omissions not captured by surface-level metrics**.

## 7 Conclusion

We introduce **DisastQA**, a large-scale, human-verified benchmark designed to evaluate model reliability under disaster-response constraints. By systematically varying evidence quality, we uncover a critical robustness gap: models excelling with perfect contexts often fail to reject noise in realistic scenarios. Furthermore, we find that surface metrics such as ROUGE overestimate factual correctness compared to keypoint-based evaluation, masking significant reliability failures. Ultimately, our results show that strong general-domain capabilities do not guarantee reliability in safety-critical settings. DisastQA thus serves as a critical stress test, shifting the paradigm from evaluating conversational fluency to ensuring verifiable, evidence-grounded robustness in high-stakes environments.

## Limitations

DisastQA currently focuses on **text-based QA in English**, prioritizing the rigorous evaluation of factual precision and evidence grounding. This scope allows us to isolate and quantify core reasoning capabilities without the confounding factors of additional modalities. While real-world disaster management involves diverse information streams, extending the framework to include *multimodal* or *multilingual* elements represents a natural direction for future research to further broaden the benchmark's applicability.

## Ethics Statement

The DISASTQA dataset is constructed from publicly available disaster-related reports intended for information sharing. All released data consist only of derived annotations (questions, answers, and keypoints), with source references provided where applicable. Automated and manual screening was applied to remove any Personally Identifiable Information (PII).

Human annotators involved in the dataset refinement were compensated fairly and informed of the nature of the task. The dataset is released solely for non-commercial research purposes. Models evaluated in this work are intended for research benchmarking and should not be used as autonomous systems in safety-critical decision-making without appropriate human oversight. In particular, erroneous or incomplete answers produced by QA systems in disaster contexts may lead to misinformed decisions, highlighting the importance of treating benchmark results as diagnostic signals rather than deployment-ready guarantees.

## References


Firoj Alam, Hassan Sajjad, Muhammad Imran, and Ferda Ofli. 2021. Crisisbench: Benchmarking crisis-related social media datasets for humanitarian information processing. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 923–932.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2024. Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17754–17762.

Xiangjue Dong, Jiaying Lu, Jianling Wang, and James Caverlee. 2023. Closed-book question generation via contrastive learning. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3150–3162, Dubrovnik, Croatia. Association for Computational Linguistics.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: Long form question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3558–3567.

Rujun Han, Peng Qi, Yuhao Zhang, Lan Liu, Juliette Burger, William Yang Wang, Zhiheng Huang, Bing Xiang, and Dan Roth. 2023. RobustQA: Benchmarking the robustness of domain adaptation for open-domain question answering. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4294–4311.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *Proceedings of the International Conference on Learning Representations*.

Muhammad Imran, Carlos Castillo, Fernando Diaz, and Sarah Vieweg. 2015. Processing social media messages in mass emergency: A survey. *ACM Computing Surveys*, 47(4).

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 2567–2577.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 6769–6781.

Daniel Khashabi, Amos Ng, Tushar Khot, Ashish Sabharwal, Hannaneh Hajishirzi, and Chris Callison-Burch. 2021. GooAQ: Open question answering with diverse answer types. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 421–433.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering

research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Bo Li, Junwei Ma, Kai Yin, Yiming Xiao, Chia-Wei Hsu, and Ali Mostafavi. 2025. Disaster management in the era of agentic ai systems: A vision for collective human-machine intelligence for augmented resilience. *arXiv*.

Ji Ma, Ivan Korotkov, Yinfei Yang, Keith B. Hall, and Ryan T. McDonald. 2021. Zero-shot neural passage retrieval via domain-targeted synthetic question generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1075–1088.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100.

Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Sarah Vieweg. 2014. Crisislex: A lexicon for collecting and filtering microblogged communications in crises. In *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media (ICWSM 2014)*, pages 376–385.

Leysia Palen and Kenneth M. Anderson. 2016. Crisis informatics—new data for extraordinary times. *Science*, 353(6296):224–225.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392.

Rajat Rawat. 2024. Disasterqa: A benchmark for assessing the performance of llms in disaster response. *arXiv*.

Koustav Rudra, Pawan Goyal, Niloy Ganguly, Muhammad Imran, and Prasenjit Mitra. 2019. Summarizing situational tweets in crisis scenarios: An extractive-abstractive approach. *IEEE Transactions on Computational Social Systems*, 6(5):981–993.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, Alex Wang, and Patrick Gallinari. 2021. QuestEval: Summarization asks for fact-based evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6594–6604.

Sarah Vieweg, Amanda L. Hughes, Kate Starbird, and Leysia Palen. 2010. Microblogging during two natural hazards events: What twitter may contribute to situational awareness. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1079–1088.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. 2024. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. In *Advances in Neural Information Processing Systems (NeurIPS 2024), Track on Datasets and Benchmarks*.

Kai Yin, Xiangjue Dong, Chengkai Liu, Lipai Huang, Yiming Xiao, Zhewei Liu, Ali Mostafavi, and James Caverlee. 2025. Disastir: A comprehensive information retrieval benchmark for disaster management. In *Findings of the Association for Computational Linguistics: EMNLP 2025*.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2024. AGIEval: A human-centric benchmark for evaluating foundation models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2299–2314.

# A Dataset Details

**Overview.** DisastQA covers eight major disaster event types: *meteorological, geophysical, hydrological, climatological, biological, technological, extraterrestrial, and conflict-induced*. Each event type includes both Multiple-Choice (MCQ) and Open-Ended (OE) questions, ensuring balanced coverage across task formats and categories. Table 4 below summarizes the distribution across all event types.

| Event Type | MCQ Count | OE Count |
|---|---|---|
| Meteorological | 250 | 125 |
| Geophysical | 250 | 125 |
| Hydrological | 250 | 125 |
| Climatological | 250 | 125 |
| Biological | 250 | 125 |
| Technological | 250 | 125 |
| Extraterrestrial | 250 | 125 |
| Conflict-induced | 250 | 125 |
| **Total** | 2000 | 1000 |

Table 4: Distribution of DisastQA questions across disaster event types. Each event type contains an equal number of questions to maintain balance across MCQ and OE tasks.

# B Human Refinement and Quality Control

This section details the human refinement process that ensured factual precision, plausibility, and clarity across all DisastQA items. Annotators systematically rewrote and verified a substantial portion of LLM-generated drafts to ensure correctness, grounding, and answerability. Each MCQ underwent independent solve-check validation by two annotators, while OE items were double-reviewed through a keypoint consensus protocol to guarantee factual completeness and consistency.

## B.1 The Necessity of Human Refinement: An Error Analysis

Although LLMs can generate scalable drafts, we observe recurring failure modes that highlight the necessity of human refinement. The most common issues include distractors lacking discriminative power (approx. 25% of MCQ drafts), factual inaccuracies (15–20%), ambiguous wording (10–15%), and incomplete factual coverage in open-ended answers (about 20–25%). These error patterns underscore that while LLMs accelerate data generation, their raw outputs fall short of the reliability standards required for safety-critical domains. Human annotators systematically corrected these issues, ensuring that all benchmark items are precise, passage-grounded, and factually complete. *Representative failure cases and a detailed breakdown of LLM failure types are provided in Appendix D.*

## B.2 Refinement Statistics

Table 5 summarizes the quantitative extent of human intervention across dataset components. The *rewrite rate* denotes the proportion of LLM drafts that were modified by human annotators due to ambiguity, weak grounding, or factual inaccuracies. For Open-Ended (OE) tasks, while LLMs produced preliminary drafts, all gold reference answers were subsequently rewritten and verified by human experts to ensure factual accuracy, conciseness, and passage fidelity, thereby eliminating residual model bias in the final dataset.

| Task Component | Rewrite Rate | QC Protocol |
|---|---|---|
| MCQ Questions | 35% | Clarity + Solve-check |
| MCQ Distractors | 45% | Plausibility validation |
| OE Questions | 30% | Clarity rewrite |
| OE Gold Answers | 80–100% | Human rewrite + factual verification |
| OE Keypoints | – | Double-pass consensus |

Table 5: Human refinement statistics across dataset components. Rewrite rate denotes the fraction of LLM drafts revised during human quality control.

## B.3 Draft–Refinement Examples

Table 6 presents representative examples of MCQ and OE refinements. Human intervention consistently improved factual precision, plausibility of distractors, and completeness of reference answers.

This section provides a full breakdown of model performance across difficulty levels. Following the keypoint-based scoring design in Section 5, we categorize Open-Ended (OE) questions into four difficulty levels—**Easy**, **Medium**, **Hard**, and **Extremely Complex**—based on the number of keypoints per reference answer. Table 7 reports Keypoint Coverage (%) across the three evaluation settings (**Base**, **Mix**, and **Golden**) for all 20 evaluated models.

These results complement the main findings in Section 6, demonstrating how evidence quality and model capacity interact with task difficulty. As shown, performance generally increases from Base → Mix → Golden for all difficulty levels, but degradation remains substantial when the number of keypoints (i.e., factual units) grows.

| Task | LLM Draft | Human Verification / Authoring |
|---|---|---|
| MCQ | Q: "What is one important precaution to follow when handling the Proflow™ Hepatitis A Test Device?" Opts: A. Store the device in direct sunlight to maintain its efficacy.; B. Ensure proper ventilation and wear personal protective equipment during use.; C. Dispose of the device in regular trash without special handling precautions.; D. Only use the device outdoors to avoid contamination. | Q: "What is one important precaution to follow when handling the Proflow™ Hepatitis A Test Device?" Opts: A. Store the device in direct sunlight to maintain its efficacy.; B. **Ensure proper ventilation and wear personal protective equipment during use.**; C. Dispose of the device in regular trash without special handling precautions.; D. Only use the device outdoors to avoid contamination. |
| MCQ | Q: "Which of the following practices should local governments implement to enhance their floodplain management under the National Flood Insurance Program?" Opts: A. Establishing a program to subsidize insurance for all property owners regardless of flood risk.; B. Implementing a mandatory evacuation plan for all residents during flood warnings.; C. Revising the floodplain ordinance to include cumulative substantial damage requirements.; D. Providing tax breaks for property development in flood-prone areas. | Q: "Which of the following practices should local governments implement to enhance their floodplain management under the National Flood Insurance Program?" Opts: A. Establishing a program to subsidize insurance for all property owners regardless of flood risk.; B. Implementing a mandatory evacuation plan for all residents during flood warnings.; C. **Revising the floodplain ordinance to include cumulative substantial damage requirements.**; D. Providing tax breaks for property development in flood-prone areas. |
| OE | *(LLM question and answer draft)* Q: "What things that people do cause acid rain to happen?" A: "The passage says acid rain is made because people burn fuels like coal and oil in factories and cars." | Q: "What are the primary human activities that lead to the formation of acid rain?" **Gold Answer (human-authored):** "The primary human activities leading to the formation of acid rain are the burning of fossil fuels in power plants and vehicles, as stated in the passage." **Keypoints (human-annotated):** {Burning of fossil fuels in power plants.; Burning of fossil fuels in vehicles.} |
| OE | *(LLM question and answer draft)* Q: "How does CBPP hurt the economy in places where it happens?" A: "CBPP makes cows sick and die, which causes big money loss for farmers and also hurts the country's economy." | Q: "How does Contagious Bovine Pleuropneumonia (CBPP) affect the economic stability of households and national economies in regions where it is prevalent?" **Gold Answer (human-authored):** "CBPP induces significant economic losses and leads to serious livestock production problems, negatively influencing people's livelihoods in affected countries." **Keypoints (human-annotated):** {CBPP causes significant economic losses.; Serious livestock production problems.; Negatively impacts people's livelihoods.} |

Table 6: Illustrative examples of MCQ and OE refinement under our Human-LLM pipeline. For **MCQ**, the LLM drafts question+options and humans refine/verify. For **OE**, the LLM drafts the question and answers; **the reference answers and keypoints are written/annotated by humans**.

## C General-Domain Comparison (MMLU-Pro Subset Evaluation)

To contextualize the domain specificity of DisastQA, we further evaluated the same set of 20 models on an **8-subject subset** of **MMLU-Pro** (Wang et al., 2024). This subset spans the subjects *Biology, Psychology, Economics, Business, Engineering, Chemistry, Law*, and *Health*, covering a total of 2,000 multiple-choice questions (each with ten options) under a *0-shot* evaluation setting.

Table 13 summarizes the average accuracies and ranks across all evaluated models. These results serve as a **general-domain baseline** for comparison with **DisastQA-MCQ (Golden)**, as visualized in Figure 1. While absolute accuracies are not directly comparable (MMLU-Pro is a closed-book benchmark, whereas DisastQA includes gold passages), the table provides a clear reference for understanding cross-domain ranking divergences discussed in Section 6.

## D Representative Error Case Analysis

To better understand the limitations of LLMs on DisastQA, we provide representative examples of common failure patterns observed across both MCQ and OE settings. This qualitative inspection complements the quantitative results in Section 6.

| Model | Easy | | | Medium | | | Hard | | | Extremely Complex | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Base | Mix | Golden | Base | Mix | Golden | Base | Mix | Golden | Base | Mix | Golden |
| GPT-5.2 | **91.03** | 94.66 | 89.22 | **86.20** | 92.33 | 95.51 | **85.89** | **96.05** | 95.23 | **100.00** | 95.00 | **100.00** |
| Gemini-3 Pro | 83.13 | 89.72 | 94.36 | 85.01 | **92.59** | **97.07** | 85.29 | 91.12 | **96.18** | 78.87 | 75.00 | **100.00** |
| GPT-4o | 82.89 | **96.22** | 96.06 | 84.42 | 91.87 | 95.25 | 80.07 | 93.06 | 95.56 | 89.58 | 90.83 | **100.00** |
| Gemini-1.5 Pro | 79.96 | 91.17 | 95.44 | 82.57 | 88.26 | 94.15 | 83.96 | 90.40 | 95.38 | 86.67 | 81.66 | 88.75 |
| Qwen-3-8B | 76.35 | 87.26 | 94.17 | 78.37 | 89.64 | 94.46 | 76.50 | 93.80 | 93.62 | 65.00 | 74.17 | 87.50 |
| Yi-6B-Chat | 67.48 | 85.71 | 86.50 | 71.72 | 84.87 | 86.23 | 76.85 | 83.04 | 88.65 | 95.83 | 96.43 | **100.00** |
| Llama-3-8B | 74.99 | 84.52 | **96.28** | 75.31 | 88.75 | 91.90 | 82.21 | 90.09 | 91.02 | 85.42 | **100.00** | 83.03 |
| Qwen-3-4B | 79.93 | 86.56 | 90.56 | 77.42 | 87.35 | 93.11 | 78.05 | 86.22 | 83.75 | 85.83 | 93.75 | 86.66 |
| Qwen-2.5-3B-Instr. | 74.98 | 85.71 | 96.17 | 76.89 | 87.54 | 91.98 | 80.33 | 87.50 | 92.71 | 81.67 | 90.83 | 81.43 |
| Phi-2 | 75.80 | 81.10 | 87.27 | 69.63 | 81.87 | 89.78 | 76.46 | 83.46 | 86.67 | 68.75 | 80.18 | 88.75 |
| Mistral-7B-Instr. | 76.00 | 85.34 | 88.88 | 75.67 | 84.59 | 89.99 | 78.82 | 81.28 | 88.27 | 82.50 | 77.08 | 90.83 |
| Gemma-7B | 66.85 | 86.36 | 93.67 | 75.23 | 83.59 | 91.53 | 75.68 | 85.32 | 93.41 | 75.42 | 77.14 | 90.83 |
| DeepSeek-v3-7B | 72.33 | 88.61 | 88.60 | 71.55 | 86.80 | 90.34 | 81.37 | 85.62 | 88.55 | 90.00 | 85.42 | **100.00** |
| Llama-3.2-3B-Instr. | 66.18 | 86.25 | 89.37 | 76.41 | 85.12 | 94.04 | 77.02 | 87.62 | 88.89 | 88.75 | 95.83 | 95.00 |
| Llama-3.2-1B-Instr. | 82.10 | 77.28 | 93.66 | 72.27 | 82.01 | 90.00 | 72.47 | 87.16 | 89.63 | 75.59 | 87.50 | 95.00 |
| Falcon-3-1B-Instr. | 77.32 | 87.33 | 91.31 | 71.90 | 84.03 | 92.61 | 74.71 | 85.29 | 91.69 | 59.58 | 83.75 | 92.86 |
| AceMath-1.5B-Instr. | 71.76 | 85.17 | 85.99 | 68.44 | 85.36 | 92.36 | 63.87 | 88.52 | 94.13 | 80.62 | 83.75 | 90.18 |
| Hunyuan-4B-Instr. | 58.79 | 85.47 | 92.50 | 66.49 | 86.13 | 89.64 | 71.81 | 90.33 | 92.63 | 95.83 | 86.66 | 81.61 |
| Hunyuan-7B-Instruct | 36.75 | 84.74 | 76.39 | 39.90 | 86.45 | 79.36 | 53.02 | 87.88 | 78.22 | 58.34 | 80.42 | 56.25 |
| Qwen-3-0.6B | 65.02 | 77.59 | 93.19 | 71.59 | 86.24 | 91.05 | 79.91 | 85.96 | 91.85 | 59.38 | 91.67 | 91.67 |

Table 7: Full breakdown of **Keypoint Coverage (%)** across four difficulty levels under **Base**, **Mix**, and **Golden** retrieval settings for all 20 models. Higher coverage indicates better factual completeness.

Errors typically fall into two categories: (1) *distractor confusion*, where models are misled by numerically or semantically similar distractors, and (2) *incomplete factual coverage*, where generated answers fail to capture all keypoints from the reference. Representative examples of these two failure types are shown in Table 8.

### D.1 Comparison of LLM Drafts and Human-Refined Versions

To further demonstrate how human verification enhances factual grounding and clarity, we provide representative cases contrasting raw LLM drafts with their human-refined counterparts. These examples illustrate the most frequent error patterns corrected during refinement— including vague question framing, factual omission, and implausible distractors. Table 9 presents four paired examples (two MCQ and two OE) showing how human experts ensured precision and passage alignment.

### D.2 Failure Mode Distribution

Table 10 summarizes the approximate distribution of common LLM draft failure types observed during the refinement phase, complementing the qualitative examples above (Table 8 and Table 9).

Finally, Table 11 details the filtering process from the DisastIR corpus to the finalized benchmark tasks, and Table 12 reports token-level statistics of annotated keypoints, which quantify factual density within the OE evaluation set.

**Keypoint Statistics.** These examples and statistics collectively confirm that while LLMs accelerate item drafting, human refinement remains indispensable for factual grounding, plausibility, and completeness in safety-critical QA benchmarks like **DisastQA**.

## E Construction Algorithm

### E.1 Filtering Process Statistics

Table 11 summarizes the stepwise filtering pipeline from the DisastIR candidate pool to the finalized DisastQA tasks. The process ensures balanced sampling across intent–event combinations, followed by human verification for both MCQ and OE items.

### E.2 Annotation Details

The following sections expand upon the annotation protocol introduced in Section 3.3 and 3.4, providing detailed task-specific guidelines for MCQ and OE verification.

#### E.2.1 General Annotation Protocol

All annotation followed a consensus-based protocol to ensure factual accuracy and consistency across the dataset.

| (a) MCQ and OE Failure Cases — Example Set 1 |
|---|
| **MCQ:** What was the reported magnitude of the earthquake in City X?<br>**Passage:** The earthquake struck City X on March 12, 2020, with a magnitude of 6.7, causing severe building damage.<br>**Options:** (A) 5.7 (B) 6.3 (C) **6.7** (D) 7.2<br>**Prediction:** (D) 7.2<br>**Error Type:** Distractor-based confusion (numerical alteration). |
| **OE:** What were the main impacts of Hurricane Y?<br>**Passage:** Hurricane Y caused widespread flooding across coastal towns, disrupted power supply to over 200,000 households, and led to multiple road closures.<br>**Reference Answer:** Flooding in coastal towns; power outage affecting 200,000 households; road closures.<br>**Model Answer:** Hurricane Y caused flooding and severe damage in coastal towns.<br>**Coverage:** 1/3 keypoints (only flooding captured; missed power outage and road closures).<br>**Error Type:** Incomplete factual coverage. |
| **(b) MCQ and OE Failure Cases — Example Set 2** |
| **MCQ:** What was the recorded peak flood depth in River Z's low-lying district?<br>**Passage:** Field reports indicate a peak flood depth of 1.5 m near the River Z levee breach, forcing evacuations in adjacent neighborhoods.<br>**Options:** (A) 0.5 m (B) 1.0 m (C) **1.5 m** (D) 2.0 m<br>**Prediction:** (B) 1.0 m<br>**Error Type:** Underestimation of numeric fact (ignored explicit value in passage). |
| **OE:** Summarize the key impacts of the Wildfire K incident.<br>**Passage:** Wildfire K burned 48,000 acres, prompted evacuations of 15,000 residents, and pushed local AQI to "hazardous" for three consecutive days.<br>**Reference Answer:** 48k acres burned; 15k residents evacuated; AQI "hazardous" for 3 days.<br>**Model Answer:** Wildfire K forced large-scale evacuations in nearby communities.<br>**Coverage:** 1/3 keypoints (mentions evacuations; misses acreage and AQI duration).<br>**Error Type:** Missing quantitative details and duration qualifiers. |

Table 8: Representative MCQ and OE failure cases from **DisastQA**. Example (a) shows numerical distractor confusion and incomplete factual coverage, while Example (b) highlights quantitative underestimation and omission of temporal keypoints—illustrating challenges in precise factual grounding despite fluent generation.

- **Annotators:** one senior Ph.D. researcher and three Ph.D. students specializing in disaster information analysis.
- **Process:** independent annotation → second-pass review → group discussion → consensus verification.
- **Division of labor:** MCQ items were verified and refined primarily by the Ph.D. students, while OE questions and reference answers were reviewed and finalized by the senior researcher.

#### E.2.2 Task-specific Annotation Guidelines for MCQ and OE

**MCQ Annotation.**

- Rewrite drafts that are ambiguous, ungrounded, or under-specified.
- Ensure that the correct answer is fully supported by the reference passage $p_g$.
- Design distractors to be plausible yet unsupported by $p_g$; avoid trivial negations or irrelevant options.

  See Table 13 for model performance on the MMLU-Pro baseline used for cross-domain comparison. This table complements Figure 1 in the main text.

**OE Annotation.**

- Rewrite drafts that are unclear or not strictly answerable from $p_g$.
- Author all reference answers manually to ensure conciseness and factual grounding.
- Decompose each reference answer into keypoints, where each keypoint represents one minimal factual unit. The number of keypoints varies with the complexity of the answer.

These keypoints form the basis for the human-verified Keypoint Coverage metric defined in Section 5.3.

**Example.** An example of human refinement illustrating conceptual improvement and distractor quality enhancement is shown in Table 14.

### E.3 Keypoint Annotation Examples

As introduced in Section 5.3, each human-authored reference answer is decomposed into minimal,

**Algorithm 1** DisastQA Construction Pipeline (Human–LLM Collaboration)

**Require:** Source corpus $\mathcal{D}$ (queries $q$, passages $p_g$)
**Ensure:** Verified MCQ set $\mathcal{Q}_{MCQ}$, Verified OE set $\mathcal{Q}_{OE}$

```
Sample (q, p_g) pairs from D based on event type
Assign each pair to MCQ or OE task
for each (q, p_g) with assigned task do
    Rewrite: LLM transforms query q into question Q using p_g
    if task = MCQ then
        LLM generates gold answer and 3 distractors
        Human verifies correctness and performs solve-check
        Add to Q_MCQ
    else                                          ▷ OE task
        LLM generates reference answer A grounded in p_g
        Human refines A for factual completeness
        if in keypoint annotation subset then
            Human annotates atomic keypoints
        end if
        Add to Q_OE
    end if
end for
```

independently verifiable factual units called *keypoints*. During evaluation, model-generated answers are scored by the proportion of gold keypoints they correctly cover, as defined by Equation 1 in the main text. Representative decompositions across multiple difficulty levels are summarized in Table 15, showing how reference answers are broken down into fine-grained factual components that capture factual completeness beyond surface overlap.

These structured examples complement the quantitative statistics in Appendix 12 and demonstrate how the keypoint-based framework operationalizes factual completeness across question complexities.

### E.4 Prompt Templates

Prompt templates follow the general generation procedure outlined in Section 5, where both Multiple-Choice (MCQ) and Open-Ended (OE) items were constructed through structured prompts grounded in retrieved passages. These templates operationalize the Base, Mix, and Golden evaluation contexts described in the main text. Prompts are provided for transparency and reproducibility, and were not tuned for individual models.

**MCQ Generation.** The following real instance demonstrates how an MCQ was generated from an actual reference passage in the **DisastQA** dataset (see Appendix H).

**OE Generation.** The OE prompt follows the same structure as the MCQ template but omits options and the correct answer, producing only the question text grounded in the passage. These templates correspond to the data generation settings used in Section 5 (*Base*, *Mix*, and *Golden* contexts).

**Mix Context Construction.** To simulate realistic noisy retrieval, each question in the Mix setting is paired with one gold passage ($p_g$) and four distractor passages ($p_d$) sampled from the DisastIR corpus across different relevance levels (scores 0–3). Specifically, we include one high-relevance distractor (score = 3 but non-gold), one medium (score = 2), one weak (score = 1), and one irrelevant (score = 0) passage whenever available. This design preserves the gold evidence while introducing both semantically similar and noisy contexts, approximating imperfect retrieval conditions where correct and misleading information coexist. These Mix contexts follow the tri-level evaluation setup described in Section 5.

## F Model Implementation Details

Table 17 summarizes all evaluated models, including model sizes, Hugging Face identifiers or API endpoints, licenses, and public access links. All open-source models were loaded via `transformers` in half-precision (fp16/bf16) mode. Closed-source models (*GPT*, *Gemini*) were accessed through their official APIs.

## G Per-model Difficulty Analysis

For completeness, we report detailed **per-model Keypoint Coverage results** across four difficulty levels (**Easy**, **Medium**, **Hard**, and **Extremely Complex**) under three retrieval settings (**Base**, **Mix**, and **Golden**). These results complement the overall difficulty trends discussed in Section 6: retrieval consistently improves factual coverage across all difficulty levels, yet performance still declines as the number of required keypoints increases.

Table 20 presents the full per-model results for all 20 evaluated models. The table is organized to facilitate side-by-side comparisons across model families, highlighting the consistent gains brought by Mix and Golden retrievals as well as the performance degradation for more complex multi-fact questions.

## H Prompt Templates

We provide the standardized prompts used for both **Multiple-Choice Questions (MCQ)** and **Open-Ended Questions (OE)** across the three evaluation settings: **Base**, **Mix**, and **Golden**.

### H.1 Multiple-Choice Questions (MCQ)

**Mix Setting (1 Golden + 4 Distractors).** This is the most challenging configuration, requiring the model to identify correct evidence among distractors.

**MCQ — Mix Prompt**
You are given 5 passages retrieved from a disaster information corpus. Some passages may be irrelevant or only partially related. Read all passages carefully and answer the question based on the most relevant and factually correct passage(s).
Write the passage used: `Passage: <1-5>`
Then write the answer: `Answer: <A-D>`
**Passage 1:** {p1}
**Passage 2:** {p2}
**Passage 3:** {p3}
**Passage 4:** {p4}
**Passage 5:** {p5}
**Question:** {q} **Options:** {opts}
Answer:

**Golden Setting (1 Golden Passage).**

**MCQ — Golden Prompt**
**Passage:** {gold_passage} **Question:** {q}
**Options:** {opts}
Answer only the letter (A–D). Answer:

**Base Setting (No Passage).**

**MCQ — Base Prompt**
**Question:** {q} **Options:** {opts}
Answer only the letter (A–D). Answer:

### H.2 Open-Ended Questions (OE)

**Mix Setting (1 Golden + 4 Distractors).** This setting evaluates the model's ability to discriminate relevant evidence from noise and generate accurate responses. Crucially, the model is explicitly instructed to identify the source passage before generating the answer, enhancing interpretability.

**OE — Mix Prompt**
You are given 5 passages (some may be irrelevant). You must select ONLY ONE passage that is most relevant to answer the question. First output the passage number you selected in the format `Passage: <single number between 1 and 5>` (only one number, no commas or multiple numbers), then provide a comprehensive answer based on that passage.
**Passage 1:** {p1}
**Passage 2:** {p2}
**Passage 3:** {p3}
**Passage 4:** {p4}
**Passage 5:** {p5}
**Question:** {query}
**Difficulty Level:** {difficulty}
**Instructions:** Provide a comprehensive answer within {word_limit} words. Ensure you cover all important aspects and key points related to the question. Be thorough but concise.
**Answer:**

***Implementation Note:*** *The* `{word_limit}` *variable is dynamically set (e.g., ≈50 words for Easy, ≈150 for Hard) to enforce information density and prevent length-based gaming of coverage metrics.*

**Golden Setting (1 Golden Passage).**

**OE — Golden Prompt**
**Passage:** {gold_passage}
**Question:** {query}
**Difficulty Level:** {difficulty}
**Instructions:** Provide a comprehensive answer within {word_limit} words. Ensure you cover all important aspects and key points related to the question. Be thorough but concise.
**Answer:**

***Note:*** *The same adaptive length constraint applies here to ensure fair comparison.*

**Base Setting (No Passage).**

> **OE — Base Prompt**
> **Question:** {query}
> **Difficulty Level:** {difficulty}
> **Instructions:** Provide a comprehensive answer within {word_limit} words. Ensure you cover all important aspects and key points related to the question. Be thorough but concise.
> **Answer:**

# I Prompt Templates and Qualitative Examples

This section provides detailed examples of the input prompts used in our evaluation, as well as qualitative examples demonstrating how model performance varies across the three settings (*Base*, *Mix*, and *Golden*).

Crucially, regarding the Open-Ended (OE) prompts, we implemented a **Complexity-Adaptive Prompting** strategy to mitigate verbosity bias. The parameter `{word_limit}` in the OE prompts is not fixed; it is dynamically adjusted based on the question's difficulty (i.e., the number of gold keypoints), enforcing concise answers for simple queries and allowing elaboration only for complex scenarios.

## I.1 Prompt Construction

Prompt templates follow the general generation procedure outlined in Section 5. Table 16 (end of section) illustrates the generation process from a raw passage to a structured MCQ.

## I.2 Qualitative Examples across Settings

To intuitively understand the difference between the settings defined in Section 5.1, we provide concrete examples in Table 18 (MCQ) and Table 19 (OE). These examples illustrate:

- **Base:** The model relies solely on internal parameters (often hallucinating).
- **Mix:** The model sees the gold passage alongside distractors (simulating noisy retrieval).
- **Golden:** The model sees only the correct passage (idealized context).

**MCQ Generation Prompt.** Table 16 demonstrates how we used LLMs to generate the initial MCQ drafts from passages.

# J Detailed Analysis by Disaster Type

To complement the aggregate results in the main text, we provide a fine-grained performance breakdown across the eight disaster categories defined in DISASTQA.

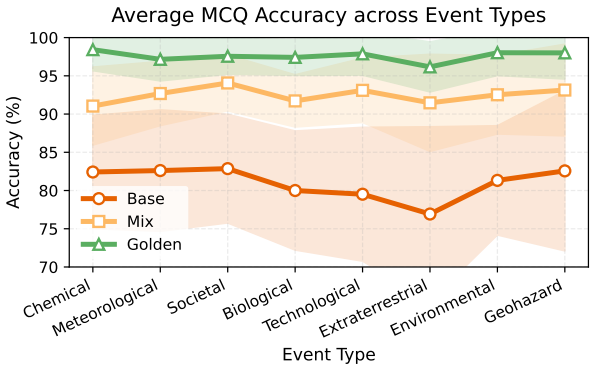


Figure 5: **MCQ Accuracy breakdown by Event Type.** The gap between *Base* and *Golden* is most pronounced in specialized domains (e.g., Biological, Extraterrestrial), confirming that models rely on retrieval for long-tail knowledge.

| LLM Draft | Human-Refined Version |
|---|---|
| **MCQ Example 1 (Medical Device Safety)** | |
| **Question:** How should you handle a medical test device?<br>**Passage:** The document is a Safety Data Sheet (SDS) for the Proflow™ Hepatitis A Test Device and Sample Preparation Device. Industries must actively engage in public campaigns to inform communities about POPs (Persistent Organic Pollutants) and their associated risks.<br>**Options:** (A) Keep it in sunlight (B) Use protection gear (C) Throw in regular trash (D) Only use outside<br>**Issue:** Question too vague; options unclear about specific safety requirements. | **Question:** What is one important precaution to follow when handling the Proflow™ Hepatitis A Test Device?<br>**Passage:** Same as left.<br>**Options:** (A) Store the device in direct sunlight to maintain its efficacy (B) **Ensure proper ventilation and wear personal protective equipment during use** (C) Dispose of the device in regular trash without special handling precautions (D) Only use the device outdoors to avoid contamination<br>**Fix:** Clarifies specific device and precise safety requirements; options clearly distinguish correct safety practices. |
| **MCQ Example 2 (Environmental Impact)** | |
| **Question:** What causes acid rain problems?<br>**Passage:** The burning of fossil fuels in power plants and vehicles contributes significantly to acid rain. This process releases sulfur dioxide and nitrogen oxides into the atmosphere, which react with water vapor to form acidic compounds.<br>**Options:** (A) Natural processes (B) Burning coal and oil (C) Water pollution (D) Industrial waste<br>**Issue:** Question too broad; options mix different pollution types; correct answer not specific enough. | **Question:** What are the primary human activities that lead to the formation of acid rain?<br>**Passage:** Same as left.<br>**Options:** (A) Natural volcanic emissions (B) **Burning fossil fuels in power plants and vehicles** (C) Agricultural fertilizer use (D) Municipal waste disposal<br>**Fix:** Specifies "human activities" and "formation"; options clearly distinguish between natural and anthropogenic sources. |
| **OE Example 1 (Electronic Waste)** | |
| **Question:** What causes electronic waste in Indian cities?<br>**Passage:** The rapid expansion of the IT industry in Bangalore is emblematic of a larger trend in many of India's urban centers. The predominant contributors to the electronic waste problem are classified into three main categories: large household appliances, which account for 42%, followed by information and communication technology (ICT) equipment at 34%, and consumer electronics making up the remaining 14%.<br>**Model Draft Answer:** IT growth causes more electronic waste and environmental problems.<br>**Issue:** Over-general; misses specific categories and percentages; lacks quantitative details. | **Question:** What are the main categories contributing to the electronic waste problem in Indian cities experiencing rapid IT industry growth, and what percentage does each category represent?<br>**Passage:** Same as left.<br>**Human Reference Answer:** The main categories are: large household appliances (42%), information and communication technology equipment (34%), and consumer electronics (14%).<br>**Fix:** Specifies exact categories with precise percentages; addresses both parts of the question comprehensively. |
| **OE Example 2 (Energy Technology)** | |
| **Question:** How can we reduce energy environmental impacts?<br>**Passage:** In light of the significant environmental impacts posed by non-combusted substances and the limitations of current combustion efficiencies, adopting cleaner and more efficient combustion technologies, alongside a transition to renewable energy sources, is imperative for lasting environmental protection.<br>**Model Draft Answer:** Use better energy methods and cleaner options to help the environment.<br>**Issue:** Over-general; misses "cleaner combustion" and "renewables transition" as explicit measures; vague phrasing. | **Question:** What concrete measures does the passage recommend to mitigate energy-related environmental impacts?<br>**Passage:** Same as left.<br>**Human Reference Answer:** The passage suggests adopting cleaner and more efficient combustion technologies and transitioning to renewable energy sources for reducing the environmental impact.<br>**Fix:** Specifies both required measures explicitly; removes vague phrasing; aligns precisely with passage content. |

Table 9: Representative MCQ and OE refinement examples from **DisastQA**. Human refinement clarifies conceptual scope, corrects factual omissions, and ensures full alignment with passage-grounded evidence.

| Failure Type | MCQ (%) | OE (%) |
|---|---|---|
| Implausible / trivial distractors | 35% | – |
| Factual inaccuracies in stems | 20% | – |
| Ambiguous wording | 15% | 20% |
| Incomplete coverage | – | 25% |
| Overly generic answers | – | 15% |
| Other minor issues | 5% | 15% |

Table 10: Distribution of common LLM draft issues observed during human refinement. Percentages are approximate, based on annotation review logs. While most drafts were structurally valid, human intervention remained essential for improving clarity, factual grounding, and coverage.

| Stage | #Items | Notes |
|---|---|---|
| Candidates (score = 3) | 5,740 | Retrieved from DisastIR |
| After stratified sampling | 3,000 | Balanced over intent × event |
| MCQ verified | 2,000 | Human-refined, solve-checked |
| OE verified | 1,000 | Human-refined, gold answers authored |
| OE keypoint subset | 200 | Double-pass consensus verified |

Table 11: Filtering process from the DisastIR corpus to the final DisastQA tasks.

| #Keypoints | Count | Mean Tokens | Min | Max |
|---|---|---|---|---|
| 1 | 5 | 15.4 | 11 | 24 |
| 2 | 104 | 38.1 | 12 | 95 |
| 3 | 181 | 65.4 | 16 | 232 |
| 4 | 244 | 97.5 | 22 | 248 |
| 5 | 259 | 123.6 | 34 | 298 |
| 6 | 129 | 144.2 | 42 | 291 |
| 7 | 52 | 156.0 | 86 | 289 |
| 8 | 15 | 182.3 | 136 | 264 |
| 9 | 6 | 215.5 | 67 | 350 |
| 10 | 3 | 188.7 | 129 | 250 |
| 11 | 0 | – | – | – |
| 12 | 1 | 209.0 | 209 | 209 |
| 13 | 0 | – | – | – |
| 14 | 1 | 240.0 | 240 | 240 |

Table 12: Answer length statistics by the number of annotated keypoints in the OE portion of DisastQA. Mean, Min, and Max are measured in tokens. Keypoint counts 11 and 13 do not occur in the dataset.

| Model | Accuracy (%) | Rank |
|---|---|---|
| GPT-5.2 | 65.2 | 1 |
| Gemini-3-Pro | 50.0 | 2 |
| Qwen-3-8B | 47.3 | 3 |
| Qwen-3-4B | 41.6 | 4 |
| GPT-4o | 41.5 | 5 |
| Gemini-1.5-Pro | 40.5 | 6 |
| Llama-3-8B | 39.9 | 7 |
| Yi-6B-Chat | 33.6 | 8 |
| Llama-3.2-3B-Instruct | 31.9 | 9 |
| Mistral-7B-Instruct-v0.2 | 31.3 | 10 |
| Hunyuan-4B-Instruct | 28.6 | 11 |
| Hunyuan-7B-Instruct | 28.3 | 12 |
| Qwen-2.5-3B-Instruct | 27.1 | 13 |
| DeepSeek-v3-7B | 25.9 | 14 |
| Gemma-7B | 25.7 | 15 |
| Phi-2 | 23.9 | 16 |
| AceMath-1.5B-Instruct | 22.5 | 17 |
| Llama-3.2-1B-Instruct | 21.0 | 18 |
| Qwen-3-0.6B | 17.3 | 19 |
| Falcon-3-1B-Instruct | 10.6 | 20 |

Table 13: Performance of 20 models on the **8-subject MMLU-Pro subset** (2,000 questions, 0-shot). This subset is used as a general-domain baseline for cross-domain ranking comparison in Figure 1.

| LLM Draft | Human Refined | Revision Rationale |
|---|---|---|
| Q: Which policy supports floodplain management? | Q: Which policy best embodies the "No Adverse Impact" principle in floodplain management? | Draft vague; refined to target the named principle and ensure grounding in the passage's definition of NAI (no increased risk, cumulative-impact review). |
| Distractor: "Approve variances when projects boost tax revenue" | Distractor: "Allow fill in floodways if compensatory storage is added offsite" | Original distractor is obviously off-policy; refined to a realistic but unsupported alternative often confused with NAI, improving plausibility without contradicting the passage. |

Table 14: Illustrative MCQ refinement demonstrating improvements in conceptual grounding and distractor plausibility for floodplain management under the "No Adverse Impact" (NAI) principle.

| Difficulty / KP | Question & Golden Answer (excerpt) | Human-Annotated Keypoints |
|---|---|---|
| **Easy (2)** | *Q:* What measures does the passage suggest for reducing the environmental impact of energy-related activities?<br>*A:* The passage suggests adopting cleaner and more efficient combustion technologies and transitioning to renewable energy sources. | **(1)** Adopting cleaner and more efficient combustion technologies.<br>**(2)** Transitioning to renewable energy sources. |
| **Medium (3)** | *Q:* What are the key responsibilities and actions that building owners must undertake after ACM cladding is identified?<br>*A:* Building owners must implement interim safety advice to ensure resident safety; measures reduce risk temporarily until remediation; and local authorities coordinate with fire services to enforce safety. | **(1)** Implement interim safety advice for resident safety.<br>**(2)** Interim measures reduce risk until remediation.<br>**(3)** Local authorities and fire services coordinate with owners. |
| **Medium (5)** | *Q:* What recent advancements have emerged in studies on suspension bridges during fire incidents?<br>*A:* Developed a numerical model using genetic algorithms to simulate cable behavior under fire; analyzed cooling with finite element methods; emphasized risks and maintenance strategies. | **(1)** Developed a numerical model for high-temperature cable behavior.<br>**(2)** Used Genetic Algorithms to optimize stiffness parameters.<br>**(3)** Highlighted hydrocarbon fire risks to bridge safety.<br>**(4)** Conducted finite element simulations for cooling.<br>**(5)** Provided recommendations for engineering practice. |
| **Hard (6)** | *Q:* How is effective decontamination achieved for all effluent within containment perimeters?<br>*A:* Achieved through validated inactivation procedures; includes pass-through autoclave, HEPA filtration, cycle alarms, gaseous decontamination, and dunk tank with active compound. | **(1)** Validated inactivation procedure.<br>**(2)** Pass-through autoclave with bioseal/interlocks.<br>**(3)** HEPA filtration of air discharge.<br>**(4)** Cycle recording/alarm systems.<br>**(5)** Gaseous decontamination chamber.<br>**(6)** Dunk tank with active compound. |
| **Extremely Complex (8)** | *Q:* How can machine learning enhance thunderstorm hazard predictions through integration of radar, satellite, and lightning data?<br>*A:* Radar data most predictive; satellite aids where radar is sparse; lightning data valuable for specific hazards; source selection affects model accuracy and overfitting. | **(1)** Integrate radar, satellite, lightning, and NWP data.<br>**(2)** Radar is the most predictive for storm indicators.<br>**(3)** Satellite imagery improves prediction where radar is sparse.<br>**(4)** Lightning data helps forecast lightning-specific hazards.<br>**(5)** Source selection impacts accuracy.<br>**(6)** Avoid overfitting via careful data selection.<br>**(7)** Understand contributions of individual data sources.<br>**(8)** Explore additional ML methods for improvement. |

Table 15: Representative keypoint annotation examples across difficulty levels. Each example shows the human-authored gold answer and its decomposition into atomic factual units (keypoints), which form the basis for factual completeness evaluation in DisastQA-OE.

**User Query:**
Enhance floodplain management under the National Flood Insurance Program (NFIP).

**Reference Passage (excerpt):**
Incorporating the ASFPM's "No Adverse Impact" policy into local floodplain management programs. Revising the floodplain ordinance to incorporate cumulative substantial damage requirements. Adopting a "no–rise" in base flood elevation clause for the flood damage prevention ordinance. Extending the freeboard requirement beyond the mapped floodplain to include equivalent land elevation. Including requirements in the local floodplain ordinance for homeowners to sign non-conversion agreements for areas below base flood elevation. Developing an educational flyer for NFIP policyholders on increased cost of compliance during post-flood damage assessments. Annually notifying owners of repetitive loss properties of Flood Mitigation Assistance funding. Offering incentives for building above the required freeboard minimum ("code plus").

**Generated MCQ:**
```
Question: Which of the following practices should local governments implement to enhance
their floodplain management under the National Flood Insurance Program?
Options:
A. Establish a program to subsidize insurance for all property owners regardless of flood
risk
B. Implement a mandatory evacuation plan for all residents during flood warnings
C. Revise the floodplain ordinance to include cumulative substantial damage requirements
D. Provide tax breaks for property development in flood-prone areas
Correct Option: C
Reason: The passage explicitly lists "revising the floodplain ordinance to include cumulative
substantial damage requirements" as a recommended practice.
```

Table 16: Real MCQ generation example from **DisastQA**. This instance shows how the generation prompt creates a question grounded in the domain passage, with realistic distractors and clear factual justification for the correct answer.

| Model | Params | Identifier / Endpoint | License | Link |
|---|---|---|---|---|
| Qwen3-0.6B | 0.75B | Qwen/Qwen3-0.6B | Apache 2.0 | HF |
| Qwen3-4B | 4.02B | Qwen/Qwen3-4B | Apache 2.0 | HF |
| Qwen3-8B | 8.19B | Qwen/Qwen3-8B | Apache 2.0 | HF |
| Llama-3.2-1B | 1.24B | meta-llama/Llama-3.2-1B-Instruct | Meta Llama 3.2 | HF |
| Llama-3.2-3B | 3.21B | meta-llama/Llama-3.2-3B-Instruct | Meta Llama 3.2 | HF |
| Llama-3-8B | 8.03B | meta-llama/Meta-Llama-3-8B-Instruct | Meta Llama 3 | HF |
| Gemma-7B | 8.54B | google/gemma-7b-it | Gemma License | HF |
| Mistral-7B | 7.24B | mistralai/Mistral-7B-Instruct-v0.2 | Apache 2.0 | HF |
| Phi-2 | 2.78B | microsoft/phi-2 | MIT | HF |
| Falcon-3-1B | 1.67B | tiiuae/Falcon3-1B-Instruct | Apache 2.0 | HF |
| Yi-6B | 6.06B | 01-ai/Yi-6B-Chat | Apache 2.0 | HF |
| DeepSeek-7B | ∼7B | deepseek-ai/deepseek-llm-7b-chat | Apache 2.0 | HF |
| Hunyuan-4B | 4.22B | tencent/Hunyuan-4B-Instruct | Apache 2.0 | HF |
| Hunyuan-7B | 7.50B | tencent/Hunyuan-7B-Instruct | Apache 2.0 | HF |
| AceMath-1.5B | 1.78B | nvidia/AceMath-1.5B-Instruct | Apache 2.0 | HF |
| GPT-5.2 | N/A | OpenAI API (gpt-5.2) | Proprietary | API |
| Gemini-3 Pro | N/A | Google API (gemini-3-pro) | Proprietary | API |
| GPT-4o | N/A | OpenAI API (gpt-4o) | Proprietary | API |
| Gemini-1.5 Pro | N/A | Google API (gemini-1.5-pro) | Proprietary | API |

Table 17: Evaluated models with parameter counts, identifiers, and license terms. "HF" denotes public Hugging Face repositories; full URLs are provided in the project's supplementary README. Note: Qwen3-0.6B corresponds to a 752M-parameter model.

| Setting | Prediction | Retrieved Passage Content (Top-1) |
|---|---|---|
| **MCQ Example 1: Volcanic Hazards** | | |
| **Question:** | *What model is utilized to enhance the accuracy of tephra dispersion predictions during volcanic eruptions?* | |
| Base | **C (Wrong)** | *No passage provided* |
| Mix | **C (Wrong)** | **[Distractor]** In: "Volcanic Hazards, Risks, and Disasters", Eds: Shroder JF... This text discusses general hazards but mimics the terminology of dispersion models... |
| Golden | **B (Correct)** | **[Gold]** Predicting Tephra Dispersion with a Mesoscale Atmospheric Model and a Particle Fall Model: Application to Cerro Negro Volcano. The study utilizes a coupled model approach... |
| **Correct Answer:** | **B** | |
| **MCQ Example 2: Shrimp Farming** | | |
| **Question:** | *What recent advancement has been made in the fight against White Spot Virus in shrimp farming?* | |
| Base | **B (Wrong)** | *No passage provided* |
| Mix | **D (Wrong)** | **[Distractor]** The company does not have the financial resources to follow this plan. In the meantime, Aquamen remains closed due to viral outbreaks... |
| Golden | **A (Correct)** | **[Gold]** Just saw a breakthrough in combating White Spot Virus! Researchers in India have developed a promising vaccine for shrimp that has shown high efficacy... |
| **Correct Answer:** | **A** | |
| **MCQ Example 3: Blowing Dust** | | |
| **Question:** | *What is the horizontal visibility range characteristic of blowing dust according to the World Meteorological Organization?* | |
| Base | **C (Wrong)** | *No passage provided* |
| Mix | **C (Wrong)** | **[Distractor]** It should be noted that due to the geographical and meteorological conditions (desert climate), El Paso may experience reduced visibility... |
| Golden | **D (Correct)** | **[Gold]** The World Meteorological Organization (WMO) categorizes dust events based on horizontal visibility: blowing dust is characterized by visibility reduction to between 1km and 10km... |
| **Correct Answer:** | **D** | |

Table 18: Qualitative analysis of MCQ failure cases. In the **Mix** setting, models are often misled by **semantically relevant distractors** (marked as [Distractor]) that share lexical overlap with the query. Only the **Golden** setting enables accurate reasoning.

| Setting | Model Response & Coverage | Context Availability |
|---|---|---|
| **OE Example 1: Laboratory Ventilation** | | |
| **Question:** | *What engineering controls and operational strategies should be implemented in laboratory ventilation systems to effectively prevent fume backflow...?* | |
| Base | To effectively prevent fume backflow... ensure safety... [General knowledge, vague details]<br>**(0% Coverage - Poor)** | *No passage provided* |
| Mix | Passage 5 states that to effectively prevent fume backflow... [Partial hallucination mixed with retrieval]<br>**(75% Coverage - Partial)** | **[Gold + 4 Distractors]**<br>*Model struggles to isolate Gold details.* |
| Golden | In laboratory environments, effective ventilation systems... [Correctly cites engineering controls]<br>**(100% Coverage - Good)** | **[Gold Only]**<br>In effect, this means arranging... |
| **Reference:** | The passage outlines several engineering controls and operational strategies for laboratory ventilation systems to prevent fume backflow and ensure safety and comfort for personnel and animals. | |
| **OE Example 2: Waste Management** | | |
| **Question:** | *What specific strategies can be implemented in public works projects to enhance waste management practices in the construction and demolition sector...?* | |
| Base | To enhance waste management... [Specific strategies focus on...]<br>**(0% Coverage - Poor)** | *No passage provided* |
| Mix | Passage 4 suggests that to enhance waste management practices... [Retrieval noise leads to omission]<br>**(67% Coverage - Partial)** | **[Gold + 4 Distractors]**<br>*Retrieval noise leads to omission.* |
| Golden | To enhance waste management practices... [Includes targeted strategies]<br>**(100% Coverage - Good)** | **[Gold Only]**<br>This can be achieved by... |
| **Reference:** | The passage outlines several strategies that can be implemented in public works projects to enhance waste management practices in the construction and demolition sector. | |
| **OE Example 3: IoT in Crowd Stampedes** | | |
| **Question:** | *How are Internet of Things (IoT) technologies being utilized in risk assessment and management strategies for early-warning systems designed to prevent crowd stampedes...?* | |
| Base | Internet of Things (IoT) technologies are increasingly being integrated...<br>**(33% Coverage - Poor)** | *No passage provided* |
| Mix | Passage 2 mentions that IoT technologies...<br>**(75% Coverage - Partial)** | **[Gold + 4 Distractors]**<br>*Partially correct but mixes sources.* |
| Golden | Internet of Things (IoT) technologies are increasingly being leveraged...<br>**(80% Coverage - Good)** | **[Gold Only]**<br>This paper discusses the analysis... |
| **Reference:** | The passage describes the utilization of Internet of Things (IoT) technologies in risk assessment and management strategies for early-warning systems aimed at preventing crowd stampedes. | |

Table 19: Progression of Open-Ended (OE) response quality. **Base** models generate fluent but factually empty responses. **Mix** settings improve coverage but often include irrelevant details from noise. **Golden** context yields the most comprehensive Keypoint Coverage.

| Part 1 Models | | | | | Part 2 Models | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Difficulty** | **Base** | **Mix** | **Golden** | **Avg** | **Difficulty** | **Base** | **Mix** | **Golden** | **Avg** |
| **AceMath-1.5B-Instr.** | | | | | **Llama-3-8B** | | | | |
| Easy | 71.76 | 85.17 | **85.99** | 80.97 | Easy | 74.99 | 84.52 | **96.28** | 85.26 |
| Medium | 68.44 | 85.36 | **92.36** | 82.05 | Medium | 75.31 | 88.75 | **91.90** | 85.32 |
| Hard | 63.87 | 88.52 | **94.13** | 82.17 | Hard | 82.21 | 90.09 | **91.02** | 87.77 |
| Extremely Complex | 80.62 | 83.75 | **90.18** | 84.85 | Extremely Complex | 85.42 | **100.00** | 83.03 | 89.48 |
| **DeepSeek-v3-7B** | | | | | **Mistral-7B-Instr.** | | | | |
| Easy | 72.33 | **88.61** | 88.60 | 83.18 | Easy | 76.00 | 85.34 | **88.88** | 83.41 |
| Medium | 71.55 | 86.80 | **90.34** | 82.90 | Medium | 75.67 | 84.59 | **89.99** | 83.42 |
| Hard | 81.37 | 85.62 | **88.55** | 85.18 | Hard | 78.82 | 81.28 | **88.27** | 82.79 |
| Extremely Complex | 90.00 | 85.42 | **100.00** | 91.81 | Extremely Complex | 82.50 | 77.08 | **90.83** | 83.47 |
| **Falcon-3-1B-Instr.** | | | | | **Phi-2** | | | | |
| Easy | 77.32 | 87.33 | **91.31** | 85.32 | Easy | 75.80 | 81.10 | **87.27** | 81.39 |
| Medium | 71.90 | 84.03 | **92.61** | 82.85 | Medium | 69.63 | 81.87 | **89.78** | 80.43 |
| Hard | 74.71 | 85.29 | **91.69** | 83.90 | Hard | 76.46 | 83.46 | **86.67** | 82.20 |
| Extremely Complex | 59.58 | 83.75 | **92.86** | 78.73 | Extremely Complex | 68.75 | 80.18 | **88.75** | 79.23 |
| **Gemini-1.5 Pro** | | | | | **Qwen-2.5-3B-Instr.** | | | | |
| Easy | 79.96 | 91.17 | **95.44** | 88.86 | Easy | 74.98 | 85.71 | **96.17** | 85.62 |
| Medium | 82.57 | 88.26 | **94.15** | 88.33 | Medium | 76.89 | 87.54 | **91.98** | 85.47 |
| Hard | 83.96 | 90.40 | **95.38** | 89.91 | Hard | 80.33 | 87.50 | **92.71** | 86.85 |
| Extremely Complex | 86.67 | 81.66 | **88.75** | 85.69 | Extremely Complex | 81.67 | **90.83** | 81.43 | 84.64 |
| **Gemma-7B** | | | | | **Qwen-3-0.6B** | | | | |
| Easy | 66.85 | 86.36 | **93.67** | 82.29 | Easy | 65.02 | 77.59 | **93.19** | 78.60 |
| Medium | 75.23 | 83.59 | **91.53** | 83.45 | Medium | 71.59 | 86.24 | **91.05** | 82.96 |
| Hard | 75.68 | 85.32 | **93.41** | 84.80 | Hard | 79.91 | 85.96 | **91.85** | 85.91 |
| Extremely Complex | 75.42 | 77.14 | **90.83** | 81.13 | Extremely Complex | 59.38 | **91.67** | **91.67** | 80.91 |
| **GPT-4o** | | | | | **Qwen-3-4B** | | | | |
| Easy | 82.89 | **96.22** | 96.06 | 91.72 | Easy | 79.93 | 86.56 | **90.56** | 85.68 |
| Medium | 84.42 | 91.87 | **95.25** | 90.51 | Medium | 77.42 | 87.35 | **93.11** | 85.96 |
| Hard | 80.07 | 93.06 | **95.56** | 89.56 | Hard | 78.05 | **86.22** | 83.75 | 82.67 |
| Extremely Complex | 89.58 | 90.83 | **100.00** | 93.47 | Extremely Complex | 85.83 | **93.75** | 86.66 | 88.75 |
| **Hunyuan-4B-Instr.** | | | | | **Qwen-3-8B** | | | | |
| Easy | 58.79 | 85.47 | **92.50** | 78.92 | Easy | 76.35 | 87.26 | **94.17** | 85.93 |
| Medium | 66.49 | 86.13 | **89.64** | 80.75 | Medium | 78.37 | 89.64 | **94.46** | 87.49 |
| Hard | 71.81 | 90.33 | **92.63** | 84.92 | Hard | 76.50 | **93.80** | 93.62 | 87.97 |
| Extremely Complex | **95.83** | 86.66 | 81.61 | 88.03 | Extremely Complex | 65.00 | 74.17 | **87.50** | 75.56 |
| **Hunyuan-7B-Instruct** | | | | | **Yi-6B-Chat** | | | | |
| Easy | 36.75 | **84.74** | 76.39 | 65.96 | Easy | 67.48 | 85.71 | **86.50** | 79.90 |
| Medium | 39.90 | **86.45** | 79.36 | 68.57 | Medium | 71.72 | 84.87 | **86.23** | 80.94 |
| Hard | 53.02 | **87.88** | 78.22 | 73.04 | Hard | 76.85 | 83.04 | **88.65** | 82.85 |
| Extremely Complex | 58.34 | **80.42** | 56.25 | 65.00 | Extremely Complex | 95.83 | 96.43 | **100.00** | 97.42 |
| **Llama-3.2-1B-Instr.** | | | | | **GPT-5.2** | | | | |
| Easy | 82.10 | 77.28 | **93.66** | 84.35 | Easy | **91.03** | 94.66 | 89.22 | 91.64 |
| Medium | 72.27 | 82.01 | **90.00** | 81.43 | Medium | 86.20 | 92.33 | **95.51** | 91.35 |
| Hard | 72.47 | 87.16 | **89.63** | 83.09 | Hard | 85.89 | **96.05** | 95.23 | 92.39 |
| Extremely Complex | 75.59 | 87.50 | **95.00** | 86.03 | Extremely Complex | **100.00** | 95.00 | **100.00** | 98.33 |
| **Llama-3.2-3B-Instr.** | | | | | **Gemini-3 Pro** | | | | |
| Easy | 66.18 | 86.25 | **89.37** | 80.60 | Easy | 83.13 | 89.72 | **94.36** | 89.07 |
| Medium | 76.41 | 85.12 | **94.04** | 85.19 | Medium | 85.01 | 92.59 | **97.07** | 91.56 |
| Hard | 77.02 | 87.62 | **88.89** | 84.51 | Hard | 85.29 | 91.12 | **96.18** | 90.86 |
| Extremely Complex | 88.75 | **95.83** | 95.00 | 93.19 | Extremely Complex | 78.87 | 75.00 | **100.00** | 84.62 |

Table 20: Coverage (%) across difficulty levels and retrieval settings (Base, Mix, Golden) for all 20 models. Each column group corresponds to one model family. The best value within each difficulty row is highlighted in bold.

| Model / Event | Chemical | Meteorological | Societal | Biological |
|---|---|---|---|---|
| **Closed-source Models** | base / mix / golden | base / mix / golden | base / mix / golden | base / mix / golden |
| GPT-5.2 | 95.2 / 95.2 / 99.6 | 92.4 / 96.4 / 99.2 | 94.8 / 97.2 / 99.6 | 90.8 / 95.6 / 100.0 |
| Gemini-3 Pro | 93.6 / 94.4 / 95.2 | 92.4 / 94.0 / 95.6 | 93.6 / 96.0 / 97.2 | 90.8 / 96.8 / 97.6 |
| GPT-4o | 93.6 / 94.0 / 99.6 | 88.8 / 96.4 / 98.8 | 89.2 / 96.4 / 99.6 | 91.6 / 94.8 / 98.4 |
| Gemini-1.5 Pro | 89.6 / 88.2 / 98.8 | 90.8 / 90.2 / 98.4 | 91.6 / 94.4 / 98.0 | 89.2 / 92.3 / 99.3 |
| **Large Open Models (6B–8B)** | | | | |
| Qwen-3-8B | 88.0 / 93.6 / 100.0 | 89.2 / 98.0 / 100.0 | 92.4 / 97.2 / 99.6 | 85.2 / 95.2 / 99.6 |
| Llama-3-8B | 86.4 / 93.6 / 100.0 | 84.4 / 95.2 / 98.0 | 87.6 / 96.4 / 98.8 | 85.6 / 95.6 / 99.2 |
| Yi-6B-Chat | 87.6 / 94.4 / 100.0 | 86.4 / 94.8 / 98.8 | 88.4 / 98.0 / 99.6 | 82.0 / 95.2 / 98.8 |
| DeepSeek-v3-7B | 81.2 / 95.2 / 100.0 | 82.4 / 94.4 / 98.4 | 85.6 / 95.6 / 98.4 | 78.8 / 92.4 / 97.6 |
| Gemma-7B | 81.2 / 86.8 / 98.4 | 85.6 / 90.8 / 98.0 | 82.4 / 92.4 / 98.4 | 78.8 / 88.8 / 97.2 |
| Mistral-7B-Instruct-v0.2 | 80.0 / 92.4 / 99.6 | 84.8 / 95.6 / 99.2 | 82.4 / 95.6 / 98.4 | 78.4 / 92.4 / 98.4 |
| Hunyuan-7B-Instruct | 64.4 / 86.8 / 93.6 | 56.8 / 87.2 / 90.4 | 64.8 / 89.2 / 93.2 | 54.0 / 88.0 / 94.8 |
| **Medium Open Models (3B–5B)** | | | | |
| Qwen-3-4B | 85.6 / 93.2 / 100.0 | 88.4 / 96.8 / 99.2 | 87.6 / 96.4 / 98.8 | 85.6 / 94.8 / 99.6 |
| Qwen-2.5-3B-Instruct | 89.2 / 95.6 / 100.0 | 87.2 / 96.0 / 99.6 | 87.6 / 96.8 / 98.4 | 81.2 / 93.6 / 98.4 |
| Llama-3.2-3B-Instruct | 86.8 / 94.8 / 99.6 | 84.8 / 95.6 / 98.8 | 83.6 / 94.8 / 97.6 | 84.0 / 94.0 / 98.8 |
| Hunyuan-4B-Instruct | 83.6 / 93.6 / 99.2 | 85.2 / 93.2 / 96.4 | 80.4 / 95.2 / 96.8 | 81.2 / 91.6 / 98.0 |
| Phi-2 | 87.6 / 95.6 / 100.0 | 89.6 / 94.8 / 96.8 | 88.4 / 96.0 / 98.4 | 85.6 / 94.0 / 99.2 |
| **Small Open Models ($\leq$2B)** | | | | |
| AceMath-1.5B-Instruct | 77.6 / 92.8 / 99.6 | 75.6 / 90.8 / 97.2 | 72.0 / 94.0 / 98.8 | 76.0 / 91.6 / 97.2 |
| Falcon-3-1B-Instruct | 77.6 / 87.2 / 98.0 | 79.6 / 92.0 / 95.6 | 75.2 / 92.0 / 97.2 | 73.6 / 89.2 / 96.4 |
| Llama-3.2-1B-Instruct | 81.6 / 92.4 / 98.4 | 83.6 / 92.8 / 98.0 | 83.6 / 94.8 / 97.2 | 79.2 / 91.2 / 95.2 |
| Qwen-3-0.6B | 79.2 / 85.2 / 88.8 | 72.8 / 84.8 / 89.2 | 74.0 / 84.8 / 89.6 | 74.4 / 84.4 / 89.6 |

*Note: Values are **Base / Mix / Golden**. Top panel shows the first four event types (Chemical, Meteorological, Societal, Biological). Bottom panel shows the remaining four (Technological, Extraterrestrial, Environmental, Geohazard).*

| Model / Event | Technological | Extraterrestrial | Environmental | Geohazard |
|---|---|---|---|---|
| **Closed-source Models** | base / mix / golden | base / mix / golden | base / mix / golden | base / mix / golden |
| GPT-5.2 | 92.4 / 96.4 / 99.6 | 89.6 / 98.0 / 99.2 | 96.0 / 96.4 / 100.0 | 94.0 / 98.4 / 100.0 |
| Gemini-3 Pro | 94.4 / 95.2 / 96.4 | 87.2 / 94.4 / 97.2 | 93.6 / 95.2 / 97.6 | 92.0 / 96.8 / 96.8 |
| GPT-4o | 90.0 / 96.8 / 100.0 | 90.8 / 98.0 / 99.2 | 90.0 / 95.6 / 99.2 | 94.4 / 98.0 / 100.0 |
| Gemini-1.5 Pro | 89.2 / 91.7 / 99.6 | 83.6 / 90.7 / 96.8 | 88.0 / 86.8 / 99.6 | 89.6 / 92.9 / 99.2 |
| **Large Open Models (6B–8B)** | | | | |
| Qwen-3-8B | 89.2 / 97.2 / 100.0 | 84.8 / 96.4 / 99.6 | 88.0 / 96.0 / 100.0 | 92.4 / 96.4 / 100.0 |
| Llama-3-8B | 86.4 / 96.4 / 100.0 | 87.2 / 94.8 / 97.6 | 88.4 / 96.0 / 99.6 | 91.2 / 97.2 / 100.0 |
| Yi-6B-Chat | 85.2 / 95.6 / 99.2 | 83.6 / 96.0 / 98.4 | 83.2 / 94.4 / 100.0 | 89.6 / 96.8 / 99.6 |
| DeepSeek-v3-7B | 79.6 / 94.0 / 99.2 | 72.0 / 92.8 / 97.6 | 80.0 / 94.8 / 99.2 | 82.0 / 95.6 / 98.8 |
| Gemma-7B | 80.0 / 91.2 / 98.0 | 71.6 / 87.2 / 97.2 | 83.6 / 88.8 / 98.4 | 78.8 / 92.0 / 98.8 |
| Mistral-7B-Instruct-v0.2 | 80.0 / 95.6 / 99.2 | 82.4 / 93.2 / 97.2 | 80.0 / 95.2 / 98.0 | 86.4 / 94.8 / 99.2 |
| Hunyuan-7B-Instruct | 54.0 / 86.8 / 93.2 | 40.4 / 86.0 / 90.0 | 65.6 / 89.2 / 94.4 | 56.0 / 90.0 / 94.8 |
| **Medium Open Models (3B–5B)** | | | | |
| Qwen-3-4B | 86.4 / 96.8 / 100.0 | 81.2 / 96.0 / 99.6 | 84.4 / 96.4 / 99.6 | 89.2 / 97.6 / 100.0 |
| Qwen-2.5-3B-Instruct | 84.4 / 96.0 / 99.6 | 81.2 / 94.8 / 98.0 | 81.6 / 94.8 / 100.0 | 90.0 / 96.0 / 99.2 |
| Llama-3.2-3B-Instruct | 81.2 / 95.2 / 98.4 | 86.4 / 96.0 / 97.6 | 84.8 / 95.2 / 99.2 | 89.2 / 96.8 / 99.2 |
| Hunyuan-4B-Instruct | 74.8 / 94.4 / 97.2 | 82.8 / 94.8 / 96.8 | 82.8 / 95.2 / 98.4 | 84.8 / 96.8 / 99.6 |
| Phi-2 | 86.8 / 97.2 / 99.6 | 83.6 / 95.2 / 98.4 | 89.2 / 96.0 / 99.2 | 87.2 / 96.8 / 100.0 |
| **Small Open Models ($\leq$2B)** | | | | |
| AceMath-1.5B-Instruct | 69.2 / 92.4 / 99.6 | 65.6 / 89.6 / 94.8 | 79.2 / 93.6 / 99.6 | 77.6 / 92.0 / 98.4 |
| Falcon-3-1B-Instruct | 74.0 / 92.4 / 96.0 | 70.8 / 91.6 / 96.4 | 74.4 / 94.0 / 98.8 | 72.8 / 91.2 / 96.8 |
| Llama-3.2-1B-Instruct | 78.0 / 92.4 / 95.2 | 76.8 / 94.0 / 95.6 | 84.8 / 95.6 / 99.2 | 89.2 / 96.8 / 99.2 |
| Qwen-3-0.6B | 70.8 / 82.4 / 88.8 | 75.2 / 77.6 / 86.4 | 72.0 / 84.0 / 87.6 | 78.0 / 80.0 / 84.8 |

Table 21: MCQ accuracy (%) under **Base / Mix / Golden** across all eight **Event Types**, shown in two panels for readability. **Top:** Chemical, Meteorological, Societal, Biological. **Bottom:** Technological, Extraterrestrial, Environmental, Geohazard. Models are grouped by access type and parameter scale.