LIFELINEQA: AN EVIDENCE-GROUNDED BENCHMARK FOR SITUATIONAL-AWARENESS QUESTION ANSWERING IN EMERGENCY OPERATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Disasters disrupt the lifeline systems that communities depend on, including safety and security, food, water, shelter, health services, energy, communications, transportation, and hazardous-material management. During response operations, decision makers need timely and reliable answers about what is happening, which lifelines are affected, where conditions are changing, and how evidence from evolving operational reports should be interpreted. Yet existing question-answering (QA) benchmarks rarely capture this setting, where information is time-sensitive, distributed across reports, and tied to domain-specific lifeline frameworks. We introduce LifelineQA, an expert-curated, evidence-grounded benchmark for lifeline-centered QA in emergency operations. LifelineQA includes three complementary question types: situational questions about incident status, threats, and response resources; lifeline-specific questions about risks, interruptions, and impacts affecting Federal Emergency Management Agency (FEMA) Community Lifelines; and complex questions requiring synthesis across reports, evidence units, or dates, including spatiotemporal reasoning. Each answerable question is linked to verifiable source evidence. We evaluate ten large language models as QA systems across five evidence conditions to assess correctness, evidence support, abstention, and temporal robustness. Results show that models answer accurately when given relevant evidence, but retrieving that evidence and citing sufficient support remain distinct barriers. LifelineQA establishes a new evaluation setting for reliable, evidence-grounded QA in high-stakes emergency operations.

1 INTRODUCTION

Disasters cause large and growing losses, which makes fast and reliable disaster response essential. Since 1980, the U.S. has experienced more than 400 weather and climate disasters that each caused at least one billion dollars in damage, with a combined cost above 2.9 trillion dollars and 16,941 deaths (NOAA NCEI, 2025). Wildfires in the Southwest have also grown more frequent, larger, and more severe (Singleton et al., 2019). The January 2025 Los Angeles wildfires show how wildfire disasters can disrupt many lifelines at once. The Palisades and Eaton fires destroyed over 16,000 structures, caused about 30 direct deaths, forced more than 200,000 evacuations, and damaged critical water, power, transportation, and communication systems (Clemens, 2025). Such events require timely and reliable information about lifeline status so emergency responders can make informed decisions. The Federal Emergency Management Agency (FEMA) groups essential services into eight Community Lifelines to report impacts and set priorities (FEMA, 2023). During disasters, accurate and timely answers about these lifelines are essential for situational awareness (Endsley, 1995; Kedia et al., 2022), helping responders identify safe work locations, determine areas requiring closure, assess resource supply, and coordinate the restoration of damaged services. This information also guides evacuations, closures, and the use of scarce resources, so a wrong or outdated answer could put lives at risk.

Responders often need to read incident reports to find the information they need. AI-enabled question answering (QA) systems could make this information easier to access by providing direct answers to specific operational questions. Recent work has increasingly applied large language models

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

(LLMs) to wildfire analysis and emergency management (Chen et al., 2025; Jiang, 2024; Xie et al., 2024), but existing QA systems leave important gaps in assessing their reliability in real-world emergency operations. First, few disaster QA benchmarks organize their questions around Community Lifelines, even though FEMA uses them to translate complex incident impacts into operational priorities (FEMA, 2023). Organizing questions by lifeline would therefore make QA evaluation more consistent with how emergency managers structure situational awareness and track the stabilization of services essential to health, safety, and economic security. Second, existing disaster QA datasets draw mainly from social media and web pages (Alam et al., 2021a; Yin et al., 2025; Chen et al., 2026) rather than the official reports that incident teams file and use, such as the Incident Status Summary (ICS-209). These reports serve as a system of record, and their contents influence each incident's priority and share of scarce resources (FEMA, 2010; NIFC, 2017), while social media posts are often brief, informal, and difficult to verify (Imran et al., 2015). Testing against official reports would show whether a system's answers hold up against the same evidence base that guides response decisions. Third, disaster QA systems must be evaluated not only on whether they retrieve supporting evidence, but also on whether that evidence is valid for the evolving situation being queried. Existing QA benchmarks often assess whether a cited passage supports an answer, but pay less attention to whether the evidence matches the relevant operational context (Dasigi et al., 2021; Gao et al., 2023; Friel et al., 2024). In emergency settings, conditions can change rapidly across reports, locations, and lifeline components; a statement that is accurate in one context may be misleading in another. QA evaluation must therefore test whether models can ground answers in evidence that is not only relevant, but also appropriate for the decision context.

To address the above challenges, we introduce LifelineQA (Figure 1), an expert-curated and evidence-grounded benchmark that provides this missing test for LLM answers about community lifelines. It contains 800 gold-standard QA pairs from ICS-209 reports on wildfire incidents in Arizona, a strong real-world test case because fires in the Southwest have grown more frequent, larger, and more severe, causing severe lifeline impacts (Singleton et al., 2019; Voiland, 2025). We designed three complementary question types to follow how responders build situational awareness, from current facts to their consequences and their change over time and across fields. Situational questions capture current facts about incident conditions and resources. Lifeline questions, guided by FEMA lifeline components and subcomponents, capture consequences for essential services by asking how they are threatened, disrupted, protected, or restored. Complex questions require synthesis across report fields or dates, including reasoning about changing conditions. We evaluate different LLMs' QA capabilities on these questions under five evidence conditions.

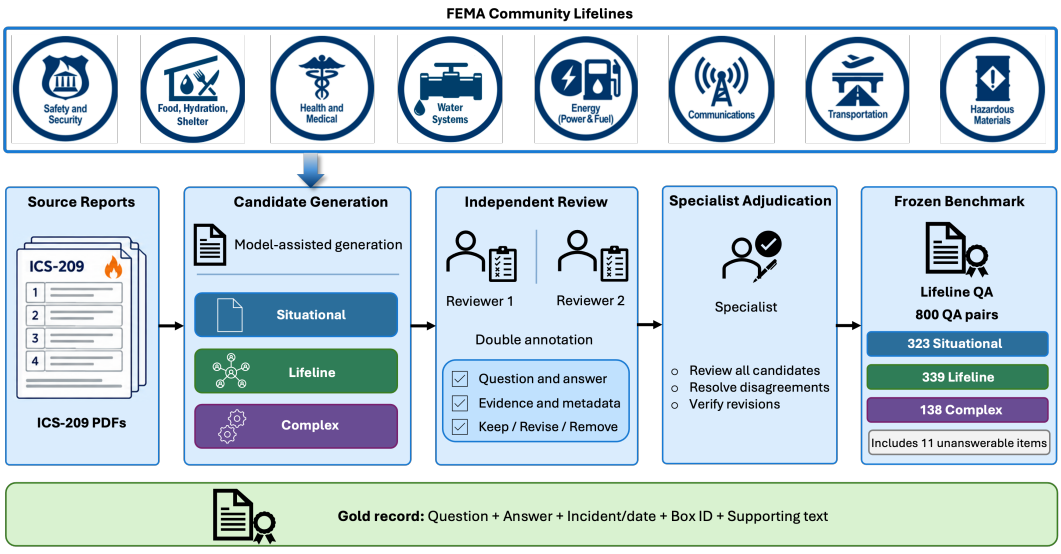

Figure 1: LifelineQA construction pipeline linking FEMA Community Lifelines, the LifelineQA question structure, and source evidence. The eight icons follow the FEMA framework (FEMA, 2023) and do not indicate equal coverage across lifelines.

In summary, LifelineQA makes three contributions. First, it connects emergency management information needs with an expert-reviewed QA benchmark over evolving official incident reports, using FEMA Community Lifelines to organize the questions. Second, each answerable item is linked to structured evidence and operational context, enabling reliable verification. Third, it provides an evaluation protocol that tests the QA capabilities of different LLMs under five evidence conditions, separating prior knowledge, reading, retrieval, contextual grounding, and answer verification.

2 Related Work

Emergency management frameworks organize incident information and guide response priorities, which makes them valuable foundations for QA benchmarks. FEMA Community Lifelines group impacts by essential service and support the reporting of status, consequences, and needs (FEMA, 2023). Most disaster QA studies rely on social media, local news, or disaster imagery (Qian et al., 2024; Al-Mohannadi et al., 2026). For example, CrisisMMD, HumAID, and CrisisBench support the classification of social media posts (Alam et al., 2018; 2021a;b); CrisisFACTS summarizes social and news streams (McCreadie & Buntain, 2023); and DisastIR and DisastQA evaluate retrieval and QA using web-based or other evidence sources (Yin et al., 2025; Chen et al., 2026).

Although useful, these sources can vary in quality and may omit critical operational information captured in official incident reports, which inform response priorities and resource allocation (FEMA, 2010; NIFC, 2017). HazardRecQA uses post-event engineering reports (Kuai et al., 2025), but its focus differs from situational awareness during an unfolding emergency. In addition, Community Lifelines have rarely been used to guide question generation in existing disaster-related QA benchmarks. To better support real-world emergency management, LifelineQA addresses these gaps by applying the framework to official incident reports and linking each answer to structured evidence and operational context for verification and fine-grained performance analysis.

Official incident reports are updated as emergencies evolve, so reliable answers must be grounded in evidence that matches the relevant operational context. Existing benchmarks address parts of this requirement. QASPER and SQuAD 2.0 evaluate evidence-based and unanswerable questions over static documents (Dasigi et al., 2021; Rajpurkar et al., 2018), while ALCE and RAGBench assess whether citations support generated answers (Gao et al., 2023; Friel et al., 2024). CRAG includes facts that change over time (Yang et al., 2024), but it does not examine repeated versions of the same operational record. This distinction matters because evidence that is accurate in one context can become misleading as conditions change. LifelineQA tests this combination of evidence grounding and contextual alignment through questions linked to structured evidence across evolving reports, including cases with competing evidence from adjacent reporting periods. Based on emergency coordination needs, we designed three question types: situational, lifeline-impact, and complex, to reflect operational needs for response and situational awareness and to evaluate QA systems across different levels of difficulty. More details can be found in Section 3.

3 The LifelineQA Benchmark

3.1 Data Source

The benchmark is built from standardized Incident Status Summary reports, also known as ICS-209 forms, which are widely used by U.S. emergency response agencies. ICS-209 forms are typically updated once per operational period, often daily. The same standardized fields may contain a single value, such as reported acreage; a short list, such as assigned resources; or several paragraphs describing significant events and projected activity over the next 12, 24, 48, and 72 hours. Because the form is standardized and repeatedly updated, each incident produces a time-ordered sequence of documents in which the same field may contain different values across reporting periods.

We generated candidate questions and answers by prompting GPT-5.6 Sol with the report PDFs. Every candidate then underwent the human review process described in Section 3.4. The resulting gold-standard QA dataset contains 800 QA pairs with detailed citation information across the three question types discussed below. Currently, the LifelineQA focuses on wildfire incidents in Arizona.

3.2 Question Types

The benchmark contains three complementary question types (Figure 2) that address emergency management information needs. Situational questions cover incident conditions and resources. Lifeline questions follow the FEMA Community Lifeline hierarchy to organize questions about essential services and community impacts. Complex questions require integrating multiple pieces of evidence as conditions evolve. Each type follows a distinct construction procedure, and every question specifies the incident and relevant operational context.

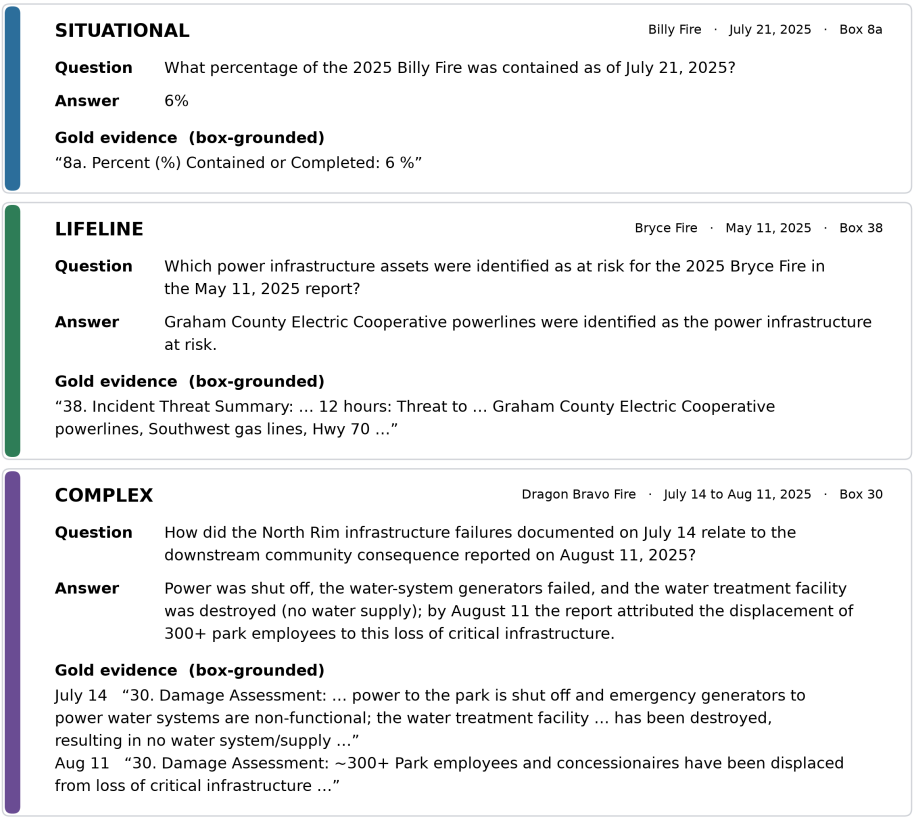

Figure 2: Examples of the three types of QA pairs presented in our benchmark.

3.2.1 Situational Questions

Situational questions cover the facts that feed a daily situation report. We co-designed a mapping schema with emergency management experts that aligns the structured and narrative fields of the report with the concepts used in operational briefings (Appendix A.1.1). Examples include incident size and containment, observed and forecasted incident behavior, threats to communities and critical assets, human impacts, assigned resources, and costs, each within the relevant operational context. These reflect questions that a duty officer may need to answer each day. Using the mapped fields, we prompted an LLM to generate clear, well-scoped questions for each incident and reporting context.

3.2.2 Lifeline Questions

FEMA Community Lifelines provide the organizing structure for the lifeline question set, guiding questions about reported threats, impacts, protective actions, and service status (FEMA, 2023). Construction began with a multi-label classification step in which an LLM identified lifeline categories and related assets, services, or disruptions mentioned in each report, such as power lines, communication repeaters, pumping stations, and road closures. For each identified lifeline category

or subcategory, the model generated factual questions tied to the incident context. These questions covered impacts, service degradation, mitigation actions, and status changes. We then applied a rule-based filter to remove candidates unsupported by the report, checking for explicit evidence and consistency with the operational context.

Table 1: The nine complex reasoning categories and the evidence span each requires.

Category	What it tests	Evidence span required	Explanation
Incident summary	Long context integration	Many fields in one report or across multiple reports	The summary of an incident is distributed across separate fields and may change over time
Comparative	Differences across fields or reports	Two or more fields, often in two reports	A comparison needs two or more values
Temporal	Timeline reasoning	The same field across different reporting periods	A trend exists across the stack of reports
Multi hop reasoning	Multi field retrieval and synthesis	Fire behavior, outlook and weather fields	The link between cause and effect crosses fields
Interdependency	Lifeline chain effects	Fields covering two or more lifelines	A chain effect crosses lifeline categories by definition
Causal	Explanation rather than fact recalls	Narrative and condition fields	The explanation is stated in one field and the effect in another
Current state reconstruction	Latest valid status of an asset or lifeline	Several reporting periods, with recency resolution	Earlier reports carry old values that need to be updated
Cascading impact reasoning	Incident to primary impact to secondary consequence for another lifeline or a community	An impact field and a consequence field across reporting periods	The second order effect is not stated directly in a single field
Early warning reasoning	At risk against actually impacted	An earlier report and a later report	The answer requires checking whether a threat materialized

LLM generation yields far more candidate QA pairs than experts can review. Annotating all candidates would overrepresent common lifeline categories and long-running incidents, while random subsampling could miss rare but important categories. We therefore frame selection as an integer linear programming (ILP) problem that chooses a limited set of QA pairs while satisfying coverage constraints across lifeline subcategories and reports. Each subcategory receives a quota proportional to its candidate count, rounded up to preserve rare valid cases. The ILP then selects candidates that meet all subcategory quotas exactly (Appendix A.1.2).

3.2.3 Complex Questions

Complex questions require integrating multiple pieces of evidence. Each complex question belongs to one of nine categories (Table 1), defined by the type of reasoning and evidence needed. Examples include questions that compare reports or evidence sources, require spatiotemporal reasoning, or involve cascading-impact reasoning. Across all categories, the gold evidence includes multiple supporting elements, either within one report or across evolving reports (Appendix A.1.3).

Categories alone do not give a balanced set, because incidents differ greatly in how much documentation they produce. Therefore, complex items are assigned to incidents in proportion to the number of reports each incident produced, and every incident receives at least one complex item regardless of how few reports it produced, usually a summary question, so the benchmark covers the full breadth of the archive.

3.3 Unanswerable Items

The gold QA includes 11 items with no supporting evidence in the available record, and their purpose is diagnostic. These are plausible operational questions whose answers are not recorded for the named incident and context. We include them because a benchmark for high-stakes reading should test whether a model invents an answer when the evidence is silent.

3.4 Expert Review and Agreement

Every QA pair in the LifelineQA gold dataset was independently reviewed by two reviewers with backgrounds in emergency management and geospatial analysis. Each reviewer recorded a yes-or-no judgment on question clarity, answer correctness, completeness of evidence support, incident and date accuracy, question-type accuracy, and validity of the unanswerable status. Each reviewer then recorded an overall decision to keep, revise, or remove the pair, along with any proposed revisions (Appendix A.1.4). A third specialist reviewed the full set, adjudicated every disagreement, and checked and edited the proposed revisions before the benchmark was finalized. The adjudicated decisions were 324 kept, 476 revised, and 4 removed.

Agreement between the two reviewers was generally substantial across the judgment dimensions (Table 2; Cohen, 1960; Landis & Koch, 1977). Question clarity requires separate treatment. It showed 98.9% raw agreement but a Cohen's κ of -0.004 because both reviewers judged nearly every question to be clear (99.8% and 99.1%). Cohen's κ can become uninformative under such prevalence imbalance (Cicchetti & Feinstein, 1990; Feinstein & Cicchetti, 1990; Byrt et al., 1993). We therefore report Gwet's AC1, which remains stable when agreement and prevalence are both high (Gwet, 2008). The AC1 for clarity is 0.989, so the low κ does not indicate unreliable review.

Table 2: Inter-annotator agreement for LifelineQA question–answer pairs.

Judgment dimension	**Raw agreement**	**Cohen's κ**
Question clarity	98.9%	-0.004
Answer correctness	85.9%	0.681
Evidence supports answer	91.7%	0.716
Incident and date correctness	97.3%	0.672
Question type correctness	89.4%	0.705
Unanswerable status validity	96.6%	0.763
Overall decision	84.3%	0.684

3.5 Freezing the Benchmark and Dataset Quality

The frozen benchmark contains 800 items (Figure 3), including 323 situational, 339 lifeline, and 138 complex items. The specialist applied the adjudicated question type labels and supplied missing gold evidence spans. The share of answerable items with nonempty gold evidence rose from 94.3% before adjudication to 100% in the frozen benchmark.

4 Evaluation Methodology

4.1 Evaluation Settings

We evaluate each model under five evidence conditions that collectively cover prior knowledge, evidence-based reading, retrieval, search within a long record, and contextual confusion. Comparisons across conditions reveal different requirements for reliable information access during emergency operations. The *Closed-Book* condition provides only the question and assesses how well a model can answer from prior knowledge. The *Gold Evidence* condition provides the annotated supporting evidence and tests performance when the required information is available. The *Retrieval* condition provides the ten highest-ranked evidence units returned by a retriever over the incident corpus. The *Full-Incident* condition provides all reports for the incident without retrieval, so the model

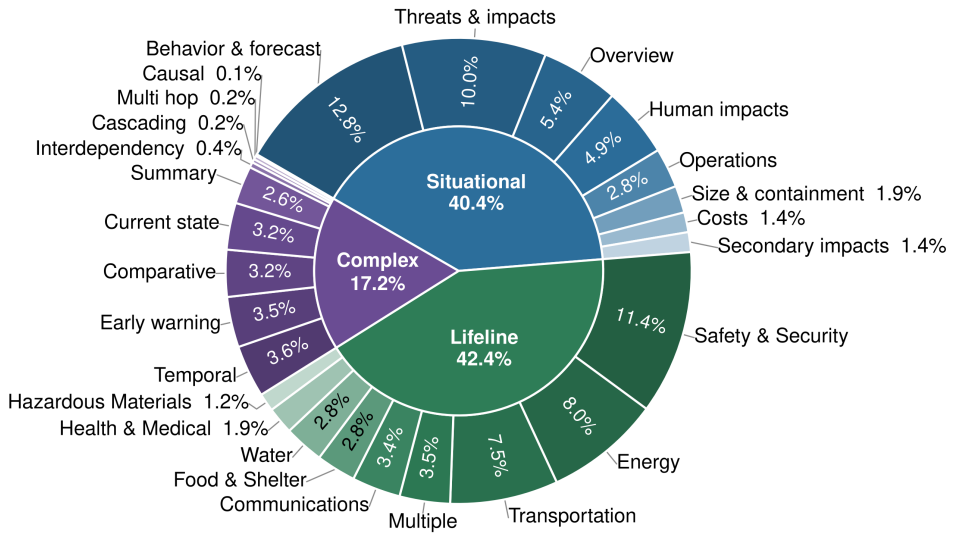

Figure 3: Distribution of questions in LifelineQA. The inner ring shows the three question families and the outer ring splits each family into its categories.

must locate the relevant evidence within the correct report. The *Gold with Temporal Distractors* condition provides the annotated supporting evidence together with corresponding evidence from adjacent reporting periods for the same incident. It retains all supporting evidence and adds only other contextual versions of that evidence, testing whether the model uses evidence that matches the event context specified in the query. Prompts are identical across all models, and the full prompt text appears in Appendix A.2.1.

4.2 Metrics

We evaluate each answer on whether it is correct and whether it can be verified. Two complementary correctness metrics measure these requirements. **Answer correctness** measures whether an answer is right. The judges (Section 4.4) grade each answer as correct, partially correct, or incorrect, scored 1, 0.5, or 0. An abstention on an answerable question also scores 0, the same as an incorrect answer. **Answer correctness** is the mean score over the answerable questions, so it reflects how good a model's answers are. **Joint answer and evidence correctness**, or joint correctness, measures whether an answer is both right and verifiable. For each answerable question, the score is 1 only if the answer is graded correct and the model cites a complete and sufficient evidence group; otherwise, the score is 0.

A paired comparison measures robustness to contextual distractors. **Temporal robustness** is the difference in answer correctness, computed by subtracting the *Gold with Temporal Distractors* condition from the *Gold Evidence* condition. It is computed on eligible questions (Table A5), where adjacent reports provide different supporting evidence, and only when a model answers in both conditions. Because the two conditions differ only by added evidence from adjacent reporting periods and shuffled order, the difference isolates confusion across evolving reports.

Additional supporting metrics explain where answers break down. **Evidence citation** metrics measure how often a model cites evidence, how well its citations overlap the annotated evidence, whether they cover the required support, and how often they cite evidence absent from the prompt. **Claim support** measures whether the statements in a model's answer are supported by the gold evidence. For each answered lifeline or complex question, we divide the response into atomic claims and use natural language inference to assess support. **Abstention** metrics measure correct refusal on unanswerable questions and unnecessary refusal on answerable questions. Appendix A.2.2 gives full definitions.

4.3 Retrieval Condition Setup

In the *Retrieval* condition, each QA model receives evidence selected from the ICS-209 reports for the incident named in the question. Retrieval operates over evidence units from the reports associated with each incident. Given a query, the retriever must locate evidence that matches the requested incident context. Each retrieval chunk contains one evidence unit and its metadata, including the incident name, report date, field number, title, and text. We compare a sparse retriever (Okapi BM25), a dense retriever (BAAI/bge-small-en-v1.5), and a hybrid retriever that combines their results through reciprocal rank fusion (Cormack et al., 2009; Robertson & Zaragoza, 2009; Xiao et al., 2024). All retrievers use standard settings specified in advance without tuning. The hybrid retriever achieves a Recall@10 of 40.2%, compared with 35.5% for the sparse retriever and 25.7% for the dense retriever. We therefore provide each QA model with the top 10 results returned by the hybrid retriever, using this cutoff as a fixed context budget. The full Recall@k curve and retrieval settings appear in Appendix A.2.3.

4.4 Evaluated Models

We evaluate ten LLMs as QA systems, including open-weight models from GPT-OSS, Gemma, Llama, and GLM, as well as proprietary Gemini models. Each model answers LifelineQA questions under the five evidence conditions, which tests its ability to provide correct answers for emergency management scenarios under different levels of information access. When evidence is provided, the model also cites the identifiers it used. This setup tests whether the model can answer within the requested incident context and identify the evidence supporting its answer. All systems receive identical prompts and condition-specific evidence; only the model endpoint differs.

We use two separate LLMs, Claude Opus 4.8 and GPT-5.6 Sol, to judge answer correctness generated by above LLMs. LLM-based evaluation is standard for open-ended answers, although it has known biases (Zheng et al., 2023). Each judge receives the question, the gold answer, and an answer without the answering model's identity. Using the same rubric, they label the answer "correct" if it conveys all essential facts, "partial" if its main point is right but incomplete or contains an error, and "incorrect" otherwise (Appendix A.2.4). These labels receive scores of 1, 0.5, and 0, respectively. An abstention on an answerable question scores 0. The two judges give the same label to 80.18% of the graded answers, with Cohen's κ of 0.465. We use their common score when they agree. A human reviewer adjudicates every disagreement.

5 Results

Table 3 shows the performance of the ten LLMs as QA systems across the five evidence conditions (see Appendix A.3 for details). Under the *Closed-Book* condition, answer correctness is near zero across all models. Under *Gold Evidence*, answer correctness rises to 0.822–0.901, which shows that the models perform well when given the annotated supporting fields. Under *Retrieval*, answer correctness declines because the required evidence may not be retrieved. These comparisons identify evidence access as a major bottleneck and motivate temporal-context-aware retrieval methods that jointly recover complementary evidence across evolving reports.

Correct answers are not always accompanied by sufficient evidence for verification. Even under the *Gold Evidence* condition, Gemma3-27B-it achieves the highest answer correctness, whereas GLM-5-2 achieves the highest joint correctness. Because answer correctness allows partial credit but joint correctness does not, the difference between the two scores is not a pure citation-error rate. Nevertheless, the different model rankings show the importance of evaluating answer content and evidence sufficiency together, especially for claims about disruptions or protective actions that decision-makers must be able to verify.

Providing the full incident history (*Full-Incident* condition) improves answer correctness but does not consistently improve joint correctness. GPT-OSS-120B and Gemma3-27B-it are excluded from this analysis because they produced malformed outputs under this evaluation setup. All eight other models that provide valid answers achieve higher answer correctness than under *Retrieval*. However, Llama4-Scout-17B shows lower joint correctness despite its improvement in answer correctness. Thus, access to more text can improve answers without ensuring sufficient evidence for verification. These results motivate methods that select and cite adequate evidence from long reporting histories.

Across question types, lifeline questions generally have the lowest answer correctness and joint answer-evidence correctness across evidence conditions, whereas complex questions achieve answer correctness comparable to the other types when evidence is provided but lower joint answer-evidence correctness than situational questions in each condition (Appendix A.3.6). This pattern highlights the difficulty of grounding lifeline and complex answers, which is often embedded in long narrative fields or distributed across multiple evidence units and reporting periods.

Compared with the *Gold Evidence* condition, the *Gold with Temporal Distractors* condition adds evidence from adjacent reporting periods. Across questions eligible for paired comparison, the resulting change in answer correctness ranges from -0.004 to 0.020, which indicates little change in answer quality (Table A5). However, joint correctness decreases for seven of the ten models across all answerable questions. Thus, answers may remain stable even when evidence grounding becomes less reliable, which highlights the need to evaluate temporal reliability through both answer correctness and supporting evidence.

Table 3: Model performance across evidence conditions. C denotes answer correctness and J denotes joint answer-evidence correctness. Bold marks the highest value and underline marks the second highest.

Model	**Closed-Book** **C**	**Retrieval** **C / J**	**Full-Incident** **C / J**	**Gold-Distractors** **C / J**	**Gold** **C / J**
Gemini-3.1-Pro-Preview	0.000	0.355 / 0.251	<u>0.809</u> / 0.560	0.861 / 0.762	0.866 / 0.769
Gemini-3.7-Flash	0.000	0.344 / 0.260	**0.815** / **0.599**	0.873 / <u>0.826</u>	0.865 / <u>0.797</u>
Gemini-3.8-Flash	0.000	0.327 / 0.260	0.731 / 0.548	0.840 / 0.797	0.843 / 0.796
Gemini-3.5-Flash-Lite	0.000	0.355 / 0.248	0.643 / 0.308	0.833 / 0.728	0.844 / 0.757
GPT-OSS-120B	0.003	0.349 / 0.191	–	0.816 / 0.638	0.824 / 0.683
Llama4-Maverick-17B	0.000	<u>0.415</u> / 0.215	0.725 / 0.299	0.847 / 0.664	0.852 / 0.719
Llama4-Scout-17B	**0.018**	0.376 / 0.223	0.634 / 0.177	0.826 / 0.659	0.822 / 0.684
Gemma3-27B-it	<u>0.011</u>	**0.492** / 0.265	–	<u>0.888</u> / 0.726	**0.901** / 0.763
Gemma4-31B-it	0.000	0.413 / <u>0.266</u>	0.774 / 0.487	**0.892** / 0.744	<u>0.891</u> / 0.771
GLM-5-2	0.000	0.381 / **0.270**	0.765 / <u>0.568</u>	0.883 / **0.842**	0.857 / **0.820**

We also conducted an exploratory *Open Web Search* condition to assess how much of the benchmark could be answered from public sources alone. Answer correctness was 0.161 for Gemini-3.1-Pro-Preview and 0.057 for GPT-OSS-120B. Because the models used different search services and some incident reports were not publicly available, we treat these results only as an external-answerability check.

6 Conclusion

LifelineQA connects FEMA Community Lifelines with situational awareness needs in emergency operations. Its questions cover lifeline impacts, incident conditions, and reasoning across evolving incident reports, with expert-reviewed evidence making each answer verifiable. Results show that evidence access is the main bottleneck. Models perform well when given gold evidence, but retrieval often misses required information. Temporal distractors further weaken evidence grounding, even when answer correctness changes less. These findings motivate retrieval methods that recover complete, relevant, and temporally aligned evidence, model training that rewards grounded answers, and evaluations that assess verifiability alongside accuracy. Operational QA systems should cite specific evidence and temporal context, and indicate when the record lacks sufficient support.

LifelineQA establishes a foundation for evaluating emergency QA across hazards, regions, and operational contexts. Current coverage reflects the source reports, so uneven question counts across three questions' subcategories and limited unanswerable questions constrain fine-grained comparisons and abstention analysis. Future work includes continuous updates, broader geographic, seasonal, and hazard coverage, a larger silver dataset, and evaluation of operational utility with emergency management practitioners.

AI USE STATEMENT

We used generative AI to generate candidate QA pairs from source PDF documents and to support automated evaluation, including LLM-based judging of model-generated answers. We also used generative AI for proofreading, grammatical editing, and readability improvements. All AI-assisted outputs were reviewed by the authors. Human experts verified all LLM-generated candidate QA pairs against the source documents before benchmark inclusion, and a human reviewer adjudicated disagreements between LLM judges. The authors take full responsibility for the final content of this work, including any text, claims, or artifacts produced with generative AI assistance.

ETHICS STATEMENT

The reports used in this work are official records of real incidents affecting real communities. The benchmark contains no personally identifiable information, but may name affected roads, facilities, and communities. It is released for evaluating reading, evidence grounding, and emergency QA systems, not for operational decision-making.

REPRODUCIBILITY STATEMENT

Sections 3–4 describe the construction pipeline and evaluation protocol, and Appendix A.2 provides prompts, metric definitions, model roster, and judge rubric. Due to privacy and security considerations, we will release a redacted processed record containing the QA evidence rather than the official ICS-209 forms. The codebase, data generation pipeline, frozen benchmark, and evaluation code will be released upon publication.

REFERENCES

A. Al-Mohannadi, A. Firoz, Y. Yang, M. Imran, and F. Ofli. DisasterVQA: A visual question answering benchmark dataset for disaster scenes. In *Proceedings of the International AAAI Conference on Web and Social Media*, 2026.

Firoj Alam, Ferda Ofli, and Muhammad Imran. CrisisMMD: Multimodal Twitter datasets from natural disasters. In *Proceedings of the 12th International AAAI Conference on Web and Social Media (ICWSM)*, 2018.

Firoj Alam, Umair Qazi, Muhammad Imran, and Ferda Ofli. HumAID: Human-annotated disaster incidents data from Twitter with deep learning benchmarks. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*, volume 15, pp. 933–942, 2021a. doi: 10.1609/icwsm.v15i1.18116.

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

Ted Byrt, Janet Bishop, and John B. Carlin. Bias, prevalence and kappa. *Journal of Clinical Epidemiology*, 46(5):423–429, 1993. doi: 10.1016/0895-4356(93)90018-V.

Yiheng Chen, Lingyao Li, Zihui Ma, Qikai Hu, Yilun Zhu, Min Deng, and Runlong Yu. Empowering LLM agents with geospatial awareness: Toward grounded reasoning for wildfire response. *arXiv preprint arXiv:2510.12061*, 2025.

Zhitong Chen, Kai Yin, Xiangjue Dong, Chengkai Liu, Xiangpeng Li, Bo Li, Junwei Ma, Yiming Xiao, Ali Mostafavi, and James Caverlee. DisastQA: A comprehensive benchmark for evaluating question answering in disaster management. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 15402–15427, 2026. URL https://aclanthology.org/2026.findings-acl.756/. arXiv:2601.03670.

Domenic V. Cicchetti and Alvan R. Feinstein. High agreement but low kappa: II. Resolving the paradoxes. *Journal of Clinical Epidemiology*, 43(6):551–558, 1990. doi: 10.1016/0895-4356(90)90159-M.

Robbie Clemens. An evaluation of the varied economic impacts of the january 2025 california wildfires. *Wesleyan Business Review*, 2025.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960. doi: 10.1177/001316446002000104.

Gordon V. Cormack, Charles L. A. Clarke, and Stefan Büttcher. Reciprocal rank fusion outperforms Condorcet and individual rank learning methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 758–759, 2009. doi: 10.1145/1571941.1572114.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, and Matt Gardner. A dataset of information-seeking questions and answers anchored in research papers. In *Proceedings of NAACL-HLT 2021*, 2021. arXiv:2105.03011.

M. R. Endsley. Toward a theory of situation awareness in dynamic systems. *Human Factors*, 37(1): 32–64, 1995.

Alvan R. Feinstein and Domenic V. Cicchetti. High agreement but low kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology*, 43(6):543–549, 1990. doi: 10.1016/0895-4356(90)90158-L.

FEMA. ICS Form 209: Incident Status Summary. NIMS ICS Forms, 2010. URL https://training.fema.gov/emiweb/is/icsresource/assets/ics%20forms/ics%20form%20209%2C%20incident%20status%20summary%20%28v3%29.pdf.

FEMA. Community Lifelines implementation toolkit, version 2.1, 2023. URL https://www.fema.gov/emergency-managers/practitioners/lifelines-toolkit.

Robert Friel, Masha Belyi, and Atindriyo Sanyal. RAGBench: Explainable benchmark for retrieval-augmented generation systems. *arXiv preprint arXiv:2407.11005*, 2024.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. Enabling large language models to generate text with citations. In *Proceedings of EMNLP 2023*, pp. 6465–6488, 2023. doi: 10.18653/v1/2023.emnlp-main.398. arXiv:2305.14627.

Kilem Li Gwet. Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, 61(1):29–48, 2008. doi: 10.1348/000711006X126600.

M. Imran, C. Castillo, F. Diaz, and S. Vieweg. Processing social media messages in mass emergency: A survey. *ACM Computing Surveys*, 47(4):1–38, 2015.

Yue Jiang. The applications of large language models in emergency management. In *2024 IEEE 6th Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC)*, volume 6, pp. 199–202, 2024. doi: 10.1109/IMCEC59810.2024.10575031.

T. Kedia, J. Ratcliff, M. O'Connor, S. Oluic, M. Rose, J. Freeman, and K. Rainwater-Lovett. Technologies enabling situational awareness during disaster response: A systematic review. *Disaster Medicine and Public Health Preparedness*, 16(1):341–359, 2022.

C. Kuai, Z. Li, B. Rosen, S. Paal, N. Jafari, J.-L. Briaud, Y. Zhang, Y. Hashash, and Y. Zhou. Knowledge-grounded agentic large language models for multi-hazard understanding from reconnaissance reports. *arXiv preprint arXiv:2511.14010*, 2025.

J. Richard Landis and Gary G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977. doi: 10.2307/2529310.

Richard McCreadie and Cody Buntain. CrisisFACTS: Building and evaluating crisis timelines. In *Proceedings of the 20th International ISCRAM Conference*, pp. 320–339, Omaha, NE, USA, 2023. doi: 10.59297/JVQZ9405.

NIFC. ICS-209 program (NIMS) user's guide, 2017. URL https://gacc.nifc.gov/oncc/docs/ICS-209_User_Guide_3.0_2017.pdf.

NOAA NCEI. U.S. billion-dollar weather and climate disasters. National Oceanic and Atmospheric Administration, 2025. URL https://www.ncei.noaa.gov/access/billions/. Retrieved September 23, 2026.

Jiale Qian, Yunyan Du, Fuyuan Liang, Jiawei Yi, Nan Wang, Wenna Tu, Sheng Huang, Tao Pei, Ting Ma, Keith Burghardt, and Kristina Lerman. Evaluating resilience of urban lifelines against flooding in China using social media data. *International Journal of Disaster Risk Reduction*, 2024. doi: 10.1016/j.ijdrr.2024.104453.

Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of ACL 2018*, pp. 784–789, 2018. doi: 10.18653/v1/P18-2124. arXiv:1806.03822.

Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, 2009. doi: 10.1561/1500000019.

Megan P. Singleton, Andrea E. Thode, Andrew J. Sánchez Meador, and Jose M. Iniguez. Increasing trends in high-severity fire in the southwestern USA from 1984 to 2015. *Forest Ecology and Management*, 433:709–719, 2019.

A. Voiland. Fires rage in Arizona. NASA Earth Observatory, 2025. URL https://science.nasa.gov/earth/earth-observatory/fires-rage-in-arizona-154541/. July 15, 2025. Retrieved September 22, 2026.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. C-Pack: Packed resources for general Chinese embeddings. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2024. arXiv:2309.07597.

Yangxinyu Xie, Bowen Jiang, Tanwi Mallick, Joshua David Bergerson, John K. Hutchison, Duane R. Verner, Jordan Branham, M. Ross Alexander, Robert B. Ross, Yan Feng, Leslie-Anne Levy, Weijie Su, and Camillo J. Taylor. WildfireGPT: Tailored large language model for wildfire analysis. *arXiv preprint arXiv:2402.07877*, 2024.

X. Yang, K. Sun, H. Xin, Y. Sun, N. Bhalla, X. Chen, S. Choudhary, R. D. Gui, Z. W. Jiang, Z. Jiang, et al. CRAG – comprehensive RAG benchmark. In *Advances in Neural Information Processing Systems*, volume 37, pp. 10470–10490, 2024.

Kai Yin, Xiangjue Dong, Chengkai Liu, Lipai Huang, Yiming Xiao, Zhewei Liu, Ali Mostafavi, and James Caverlee. DisastIR: A comprehensive information retrieval benchmark for disaster management. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 1836–1867, 2025. URL https://aclanthology.org/2025.findings-emnlp.97/. arXiv:2505.15856.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems 36, Datasets and Benchmarks Track*, 2023. arXiv:2306.05685.

A Appendix

A.1 Benchmark Construction Details

A.1.1 Situational Questions: Field-to-Concept Mapping and Generation Prompt

Situational questions target the facts that feed a daily situation report. We co-designed a mapping schema with emergency-management experts that aligns the structured and narrative fields of the ICS-209 form with the concepts used in operational briefings (Table A1). The schema groups form content into eight concept families, because this is how incident information is consumed under time pressure. Given the mapped fields for one incident and report date, we prompt an LLM to write one

clear, well-scoped question whose answer is drawn from that report; every question is required to name the incident, the year, and the report date.

Generation prompt. The prompt embeds a block of review rules co-designed with the domain experts and constrains every evidence citation to a fixed whitelist of ICS-209 box identifiers. No gold evidence snippet or expected label is shown to the model, so it must locate its own evidence; unanswerable items receive the identical prompt.

```
You are an expert annotator for wildfire incident reports (ICS-209). Your job is to create
     one high-quality Situational QA pair.

TASK: Write one question that captures this operational intent, answered from the provided
     ICS-209 {source}:
    "{intent} for the {incident_display} Fire on {date_display}?"
The document filename is "{pdf_filename}".

[Expert review rules embedded here --- see below]

REQUIREMENTS
1. Question - rewrite the intent into clear, natural language an emergency manager would use
     ; it MUST contain the incident name, the year, and the report date; simple structure,
     no nested clauses.
2. Answer  - extract the gold answer exactly from the {source}; name the specific entities (
     assets, sites, roads, communities) where the report names them.
3. Evidence - doc_id must be exactly "{pdf_filename}"; box_id MUST be one of the valid ids
     listed below (do not invent, abbreviate, or merge boxes); the intent usually maps to: {
     pref}; evidence_text must be an exact quote from that box.
4. Metadata - question_type: "factual"; difficulty: "easy"|"medium"|"hard"; id: "{task_id}_{
     theme}".

JSON OUTPUT FORMAT (strict)
{
  "id": "{task_id}_{theme}",
  "incident_name": ["<YEAR> <INCIDENT NAME>"],
  "incident_number": ["<INCIDENT NUMBER>"],
  "question": "<final rewritten natural language question>",
  "question_type": "factual",
  "is_impossible": false,
  "tags": ["tag1", "tag2"],
  "gold_answer": "<short authoritative answer naming the entities>",
  "gold_evidence": [
    {"doc_id": "{pdf_filename}", "box_id": "<one of the valid ids>",
     "evidence_text": "<exact quote from that box>"}
  ]
}
If no evidence exists: is_impossible=true, gold_answer="", gold_evidence=[].
OUTPUT: Return EXACTLY ONE JSON object. No commentary.
Expert review rules embedded in the situational (and lifeline) prompts.
1.  Prioritise operationally useful questions about named entities, communities, critical
     infrastructure, human impacts, access, protective actions, and operational needs.
2.  Match the question's specificity to the evidence. If you ask "which named road / area /
     asset?", the gold answer MUST name it. Repeating the category word from the question is
      not an answer.
3.  If the report affirmatively records only a generic category or a checked status, do NOT
     set is_impossible; rewrite as a category/status question ("Were any road closures in
     effect...?"). Set is_impossible=true ONLY when the intent cannot be answered even after
      such a rewrite.
4.  Distinguish carefully among threatened / potentially threatened / decreased threat /
     impacted /damaged / destroyed / closed / operational / protected. Never convert one
     status into another.
5.  "Decreased threat" does NOT mean "currently at risk".
6.  Missing information is NOT "no". If absent, is_impossible=true. If the report explicitly
      states zero / no threat / no impact, that IS answerable.
7.  The question date, doc_id, gold answer, and evidence must all refer to the SAME report.
8.  Evidence must directly support the question's status, time horizon, and subject; select
     evidence by MEANING regardless of which box it sits in; a planned action is not proof
     of current impact.
9.  Use the smallest exact evidence that supports every statement in the answer.
10. Avoid redundant questions whose answer is already in the question.
11. Avoid vague or artificial phrasing an emergency manager would not use.
12. Ask ONE operational intent per question.
13. Keep this a SITUATIONAL question: a single-report factual lookup (not a temporal
     comparison, not a lifeline severity assessment).
14. Do NOT infer counts, costs, closures, impacts, or downstream risks not explicitly stated
      (the fields reviewers most often caught as hallucinated).
```

List of the valid ICS-209 box identifiers shared by all generation prompts:

```
box_1_incident_name, box_2_incident_number, box_3_report_version, box_4_incident_commander,
box_5_incident_management_organization, box_6_incident_start_date_time, box_7_incident_size,
box_8a_percent_contained, box_8b_percent_perimeter_to_be_contained, box_9_incident_type,
box_9b_incident_description, box_9c_cause, box_9d_fire_suppression_strategy,
box_10_incident_complexity_level, box_11_report_time_period, box_14_date_time_submitted,
box_15_primary_location_organization_agency_sent_to, box_16_state,
    box_17_county_parish_borough,
box_18_city, box_19_unit_or_other, box_20_incident_jurisdiction,
    box_21_incident_location_ownership,
box_22_latitude_longitude, box_23_us_national_grid_reference, box_24_legal_description,
box_25_short_location_description, box_26_utm_coordinates, box_27_geospatial_data,
box_28_observed_fire_behavior, box_29_primary_fuel_model, box_30_damage_assessment,
box_30_damage_assessment_structural_summary, box_31_public_status_summary,
box_32_responder_status_summary, box_33_life_safety_health_status_threat_remarks,
box_34_life_safety_health_threat_management, box_35_weather_concerns,
box_36_projected_incident_activity, box_37_strategic_objectives,
    box_38_incident_threat_summary,
box_39_critical_resource_needs, box_40_strategic_discussion, box_41_planned_actions,
box_42_projected_final_incident_size, box_43_anticipated_containment_date,
box_44_projected_demobilization_start, box_45_estimated_costs_to_date,
box_46_projected_final_incident_cost, box_47_remarks, box_50_overhead,
    box_51_total_personnel,
box_52_total_resources, box_53_additional_cooperators
```

Table A1: Mapping schema between ICS-209 fields and situational reporting concepts.

Concept family	**Situational concepts**
Incident overview	Incident number; report version; location; ignition or start date and time; incident type
Size and containment	Current incident size in acres; percent contained
Incident behavior and forecast	Observed incident behavior; 12, 24, 48 and 72 hour outlooks; longer term outlook; weather concerns
Threats and impacts	Evacuation status; road closures; area closures; critical assets at risk
Human impacts	Civilian fatalities; civilian injuries; number evacuated or displaced; sheltering in place; temporary shelters; public health concerns
Operations and resources	Incident management team type; total personnel assigned; critical resource needs at 12, 24, 48 and 72 hours
Secondary and cascading impacts	Cascading impacts or secondary risks; ecological follow-on risks
Costs and projections	Estimated costs to date; projected final cost estimate

A.1.2 Lifeline Questions: Subcomponent Taxonomy, Selection ILP, and Prompts

Lifeline questions cover impacts to the FEMA Community Lifelines and their subcomponents. Construction has three stages: (i) an LLM labels which lifeline subcomponents each report meaningfully mentions, against the full FEMA taxonomy (Table A2); (ii) for each labelled subcomponent the model writes short, operational questions tied to that incident and report date, after which a rule-based filter removes candidates the report does not support; and (iii) because this yields far more candidates than experts can review, an integer linear program (ILP) selects a compact, coverage-preserving review set.

Table A2: FEMA Community Lifelines taxonomy embedded in the subcomponent-classification prompt.

Lifeline	**Component**	**Subcomponents**
Safety and Security	Law Enforcement/Security	Police Stations; Law Enforcement; Site Security; Correctional Facilities
	Fire Service	Fire Stations; Firefighting Resources
	Search and Rescue	Local Search and Rescue

Lifeline	Component	Subcomponents
	Government Service	Emergency Operation Centers; Essential Government Functions; Government Offices; Schools; Public Records; Historic/Cultural Resources
	Community Safety	Flood Control; Other Hazards; Protective Actions
Food, Hydration, Shelter	Food	Commercial Food Distribution; Commercial Food Supply Chain; Food Distribution Programs
	Hydration	Temporary Hydration Missions; Commercial Water Supply Chain
	Shelter	Housing; Commercial Facilities (e.g., hotels)
	Agriculture	Animals and Agriculture
Health and Medical	Medical Care	Hospitals; Dialysis; Pharmacies; Long-Term Care; VA Health System; Veterinary Services; Home Care
	Patient Movement	Emergency Medical Services
	Fatality Management	Mortuary and Post-Mortuary Services
	Public Health	Health Surveillance; Human Services; Behavioral Health; Vector Control; Labs
	Medical Supply Chain	Blood/Blood Products; Manufacturing; Pharmaceutical Devices; Medical Gases; Distribution; Critical Clinical Research; Sterilization; Raw Materials
Energy	Power Grid	Generation Systems; Transmission Systems; Distribution Systems
	Fuel	Refineries/Fuel Processing; Fuel Storage; Pipelines; Fuel Distribution; Off-shore Oil Platforms
Communications	Infrastructure	Wireless; Cable/Wireline; Broadcast (TV and Radio); Satellite; Data Centers/Internet
	Alerts, Warnings, Messages	Local Alert/Warning Ability; Access to IPAWS (WEA, EAS, NWR); NAWAS Terminals
	911 & Dispatch	Public Safety Answering Points (PSAP); Dispatch
	Responder Communications	Land Mobile Radio (LMR) Networks
	Finance	Banking Services; Electronic Payment Processing
Transportation	Highway/Roadway/ Motor Vehicle	Roads; Bridges
	Mass Transit	Bus; Rail; Ferry
	Railway	Freight; Passenger
	Aviation	Commercial; General; Military
	Maritime	Waterways; Ports and Port Facilities
Hazardous Materials	Facilities	Oil/HAZMAT Facilities; Oil/HAZMAT/Toxic Incidents from Facilities
	HAZMAT, Pollutants, Contaminants	Oil/HAZMAT/Toxic Incidents from Non-Fixed Facilities; Radiological or Nuclear Incidents
Water Systems	Potable Water Infrastructure	Intake; Treatment; Storage; Distribution
	Wastewater Management	Collection; Storage; Treatment; Discharge

Selection integer linear program. Let $\mathcal{F}$ denote the set of documents (reports) and $\mathcal{S}$ the set of lifeline subcomponents, with each document d associated to an incident $f(d)$. A binary availability indicator $A(d, s)$ equals 1 if document d contains sufficient information to generate a QA pair for subcomponent s. For each subcomponent we count candidate questions across the corpus to obtain a global frequency $C(s)$, and set the quota:

$$Q(s) = \lceil \beta \cdot C(s) \rceil, \qquad \beta \in (0, 1).$$

The ceiling guarantees that a subcomponent with rare but nonzero evidence survives compression. Binary decision variables $x(d, s)$ indicate whether a QA pair for subcomponent s is drawn from document d; auxiliary variables $y(d)$ and $z(f)$ mark documents and incidents counted in the objective. The program is:

$$\max \sum_{d \in \mathcal{F}} y(d) \; + \; \lambda \sum_{f \in \mathcal{E}} z(f), \qquad \lambda \ll 1,$$

subject to

$x(d,s) \le A(d,s)$	draw only from documents that support the pair
$\sum_d x(d,s) = Q(s)$	meet each subcomponent quota exactly
$x(d,s) \le y(d) \le z(f(d))$	used-document / used-incident linking
$z(f) \le \sum_{d:\, f(d)=f} y(d)$	an incident counts only if one of its documents is marked
$x(d,s),\ y(d),\ z(f) \in \{0,1\}$	binary decision variables

The objective is intended to favor distinct documents and, through the small secondary term ($\lambda \ll 1$), distinct incidents.

Generation prompt. The subcomponent-classification prompt reads a report PDF and returns the meaningfully mentioned subcomponents from Table A2 as JSON. The question-generation prompt then produces one operational question per subcomponent, sharing the box whitelist and a lifeline specific rule block:

```
You are an expert annotator for wildfire incident reports (ICS-209). Create one high-quality
      LIFELINE QA pair for evaluating a RAG system.

TASK: Write one operational question capturing this intent, answered from the provided ICS
     -209 {source}:
    "{intent} for the {incident_display} Fire on {date_display}?"
This concerns the FEMA "{lifeline}" community lifeline. Choose the most specific matching
     subcomponent(s) from: {subs}. The document filename is "{pdf_filename}".

[Expert lifeline rules embedded here --- see below]

REQUIREMENTS
1. Question: operational (risk/impact/status/action/consequence), natural wording, contains
     the incident name, year, and report date; one intent only.
2. Answer: extract exactly from the {source}; name the specific asset/route/facility/
     community.
3. Evidence: doc_id = "{pdf_filename}"; box_id MUST be one of the valid ids; usually in {
     pref}; evidence_text must be an exact quote.
4. Metadata: question_type "lifeline"; lifeline_subcomponent from {subs}; difficulty; id.

JSON OUTPUT FORMAT (strict)   -- identical to the situational schema plus a "
     lifeline_subcomponent": ["<one of the listed subcomponents>"] field.
If no evidence exists: is_impossible=true, gold_answer="", gold_evidence=[].
OUTPUT: Return EXACTLY ONE JSON object. No commentary.
EXPERT LIFELINE RULES
1. Ask OPERATIONALLY, never as a taxonomy dump: risk / actual impact / current status /
     protective actions taken / service or infrastructure consequence. Not "What information
      was reported about X?".
2. Name the specific asset, route, facility, or community where the report names it.
3. Distinguish status exactly: threatened / potentially threatened / decreased threat /
     impacted / damaged / destroyed / disrupted / closed / operational / protected /
     restored. Never convert one into another. "Decreased threat" is NOT "currently at risk
     ".
4. Missing information is NOT "no". Absent -> is_impossible=true; explicit "no impact/no
     threat" IS answerable.
5. Evidence must directly support the question's status, time horizon, and subject; select
     by MEANING.
6. Do NOT infer counts, costs, or downstream effects not stated.
7. A cross-lifeline question must stay a SINGLE-report factual lookup with a directly-stated
      consequence; if it needs multi-step reasoning or multiple dates, it is Complex.
8. Keep it a LIFELINE question: a single-report operational fact about a FEMA community
     lifeline.
```

A.1.3 Complex Questions: Worked Examples, and Prompt

One worked example per category from the frozen gold (evidence trimmed where long, shown by . . .). Cross-report evidence lists one {report :: box} pair per line.

```
1. Incident summary
Q: What was the overall incident status and key conditions reported for the 2025 Coyote Fire
      on May 11, 2025?
A: The Coyote Fire was 1,000 acres in size with 0% containment, exhibited active fire
     behavior, and was burning in brush, short grass, tall grass, juniper, and pinyon pine
     near Springerville, Arizona.
Evidence: Coyote 5.11.2025 :: box_7_incident_size + box_8a_percent_contained +
     box_28_observed_fire_behavior + box_29_primary_fuel_model +
     box_25_short_location_description (5 boxes, one report)
```

```
2. Comparative
Q: Compare the containment percentage reported for the White Sage Fire on July 16, 2025 and
     July 24, 2025.
A: On July 16, 2025 containment was 9%; on July 24, 2025 containment was 77%.
Evidence: White Sage 7.16.2025 :: box_8a_percent_contained  |  White Sage 7.24.2025 ::
     box_8a_percent_contained
3. Temporal
Q: How did the reported fire size progress over time between July 14, 2025 and August 20,
     2025 for the 2025 Billy Fire?
A: The fire grew from 752 acres on July 14, 2025 to 27,522 acres by August 20, 2025.
Evidence: Billy 7.14.2025 :: box_7_incident_size  |  Billy 8.20.2025 :: box_7_incident_size
4. Multi-hop reasoning
Q: Using fuel models, weather concerns, and projected incident activity, explain why single-
     tree torching remained a risk for the 2025 Bronco Fire during August 11-13, 2025.
A: Grass-dominated fuels (models 101, 122, 147) with intermixed mature juniper and a monsoon
     -stressed pinyon/juniper overstory left stressed trees available to torch; moderate
     winds and continued hot, dry weather through Aug 13 kept single-tree torching possible,
      with unpredictable outflow winds able to move fire on either flank absent rainfall.
Evidence: Bronco 8.11.2025 :: box_29_primary_fuel_model + box_35_weather_concerns  |  Bronco
      8.13.2025 :: box_35_weather_concerns + box_36_projected_incident_activity
5. Interdependency
Q: For the 2025 Dragon Bravo incident, how did the North Rim infrastructure failures
     documented on July 14 relate to the downstream community consequence explicitly
     reported on August 11, 2025?
A: July 14 documented park power shut off, non-functional water-system generators, and a
     destroyed water treatment facility (no water supply); by August 11 the report
     attributed the displacement of 300+ park employees and concessionaires to that loss of
     critical infrastructure.
Evidence: Dragon Bravo 7.14.2025 :: box_30_damage_assessment  |  Dragon Bravo 8.11.2025 ::
     box_30_damage_assessment
6. Causal
Q: How did measurable precipitation cause changes in observed fire behavior for the 2025
     Bronco Fire between August 13, 2025 and August 14, 2025?
A: Measurable rain across most of the fire reduced activity from moderate to minimal;
     backing, creeping and smoldering became dominant and the potential for single-tree
     torching was significantly reduced.
Evidence: Bronco 8.13.2025 :: box_28_observed_fire_behavior  |  Bronco 8.14.2025 ::
     box_28_observed_fire_behavior + box_36_projected_incident_activity
7. Current-state reconstruction
Q: For the 2025 Cody incident, how did the status of the Zone 18 evacuation order in the
     Oracle area change from the May 26, 2025 report to its latest mention on May 27, 2025?
A: Zone 18 (Oracle area) was under a Go-Level 3 evacuation order on May 26; in the latest
     mention, May 27, the order was scheduled to be lifted as of 1900 that day.
Evidence: Cody 5.26.2025 :: box_33_life_safety_health_status_threat_remarks  |  Cody
     5.27.2025 :: box_33_life_safety_health_status_threat_remarks
8. Cascading impact reasoning
Q: During the 2025 White Sage incident, how did the U.S. Highway 89A and fire-area access
     closures documented on July 16 lead to a secondary consequence for adjacent communities
      reported on July 17, 2025?
A: On July 16, U.S. Highway 89A was closed southbound at SR 389 and northbound at U.S. 89,
     and closure orders restricted public access to the fire area; by July 17 the report
     named long-term economic impacts in adjacent communities due to prolonged road and area
      closures, as well as mandatory evacuations.
Evidence: White Sage 7.16.2025 :: box_33_life_safety_health_status_threat_remarks  |  White
     Sage 7.17.2025 :: box_38_incident_threat_summary
9. Early-warning reasoning
Q: For the 2025 Cody incident, did the threat to mixed commercial/residential properties
     reported on May 22, 2025 materialize as damage or destruction by the May 27, 2025
     report?
A: No. On May 22, 150 mixed commercial/residential properties were threatened with none
     damaged or destroyed; by May 27 that category showed 0 threatened, 0 damaged, 0
     destroyed.
Evidence: Cody 5.22.2025 :: box_30_damage_assessment_structural_summary  |  Cody 5.27.2025
     :: box_30_damage_assessment_structural_summary
```

Generation prompt. All of one incident's dated reports are assembled and prepended to the prompt. A shared template carries a category slot; category-specific reasoning shapes are injected from a per-type rule block. The shared rules require multi-report grounding:

```
You are an expert annotator for wildfire incident reports (ICS-209). You are given MULTIPLE
     dated reports for ONE incident (prepended above). Create ONE high-quality COMPLEX QA
     pair of the type below.

INCIDENT: {incident_display}  (reports span {date_from} to {date_to})
COMPLEX TYPE: {category}
WHAT IT SHOULD TEST: {intent}
HOW TO BUILD IT: {guidance}
The incident's reports (cite these EXACT filenames as doc_id): {files}
```

```
COMPLEX QA RULES
1. MULTI-EVIDENCE REQUIRED: the question MUST require reasoning across >= 2 DISTINCT boxes.
     CROSS-REPORT is additionally REQUIRED for Temporal, Current-State, Early-Warning, Event
      Summary, and any Comparative that compares across dates (each must reason across >= 2
     report dates). Multi-hop, Causal and Interdependency may draw their >= 2 boxes from one
      report OR across reports (prefer across reports when the reasoning spans dates).
2. gold_evidence MUST cite the >= 2 DISTINCT boxes the answer depends on; each item is {
     doc_id, box_id, evidence_text} with evidence_text an EXACT quote. For the cross-report
     types above, those boxes MUST come from >= 2 DISTINCT reports (>= 2 doc_id).
3. The question must name the incident, the year, and the relevant date(s).
4. Ground the answer ONLY in the reports; do not infer facts not stated; preserve status
     words exactly (at risk / threatened / impacted / damaged / destroyed / closed /
     evacuated / protected / repopulated / restored).
5. If the reports genuinely do not support this type, set is_impossible=true.

[type-specific rule for {ctype} injected here --- one of the nine below]
JSON OUTPUT FORMAT (strict): question_type "complex", complex_categories ["{category}"], a
     type-specific "analysis" object, and gold_evidence citing the required boxes/reports.
     OUTPUT: one JSON.
TYPE RULES (one injected per item)
- Current-State Reconstruction: track ONE named asset/lifeline; cite an EARLIER report for
     its prior status and the LATEST report that mentions it for its current status;
     analysis = {tracked_entity, prior_status, current_status, prior_doc_id, current_doc_id
     }.
- Cascading Impact: trace a primary impact -> a secondary consequence (same or different
     reports); the causal link must be shown by causal language in the evidence or a later
     back-reference; analysis = {primary_impact, secondary_consequence, causal_link_text}.
- Early-Warning: prior_doc_id supports ONLY a risk/threat; outcome_doc_id (later) sets
     outcome_status in {materialized, protected, not_materialized, unresolved}; analysis = {
     tracked_entity, prior_status, outcome_status, prior_doc_id, outcome_doc_id}.
- Event Summary: summarize the WHOLE incident; phases_covered lists >= 2 distinct phases and
      evidence cites reports from more than one phase; analysis = {phases_covered:[..]}.
- Comparative: compare >= 2 explicit values of the same/related fields (two boxes of one
     report, or the same box across two dates); state every value and the direction/size of
     the difference; analysis = {compared_field, values:[{doc_id, value}], difference}.
- Temporal: track the SAME field/box across >= 2 report dates; describe the trend/timeline
     and name the dates; report only stated values (do not interpolate); analysis = {
     tracked_field, timeline:[{doc_id, value}], trend}.
- Multi-hop: synthesize >= 2 DIFFERENT fields (e.g. fire behaviour + weather + projected
     activity) whose combination no single box states; lay out each linked fact and the step
      joining them; analysis = {reasoning_chain:[{doc_id, box, fact}], conclusion}.
- Interdependency: trace a chain effect crossing >= 2 FEMA lifelines (crossing lifelines is
     the defining test, vs Cascading); the inter-lifeline link must be stated or
     referentially implied; analysis = {lifelines:[..], chain_effect, doc_ids:[..]}.
- Causal: explain WHY -- the cause stated in one box, the effect in another; give cause,
     effect, and the causal statement linking them, using only causal language present in
     the evidence; analysis = {cause, effect, causal_link_text, cause_doc_id, effect_doc_id
     }.
```

A.1.4 Annotation Instructions for Expert Review

Each of the 804 candidate items was annotated independently by two reviewers with emergency-management and geospatial backgrounds, using an offline review tool that shows the candidate QA beside the source report PDF. For each item a reviewer recorded a yes/no judgment on six dimensions and then an overall decision with an optional revision. A third specialist then reviewed the full set, adjudicated every disagreement, and edited the proposed revisions before the benchmark was frozen. The instructions given to reviewers were:

- **Question clarity:** is the question clear, unambiguous, well-scoped, and phrased as an emergency manager would ask it? It must name the incident, the year, and the report date.
- **Answer correctness:** is the gold answer factually correct against the named report, with the right entities, numbers, dates, and polarity/negation?
- **Evidence supports answer:** does the cited evidence (doc_id, box_id, exact quote) fully support the answer? The quote must be verbatim from that box and sufficient on its own; multiple boxes are allowed when the answer spans fields.
- **Incident and date correctness:** do the question, the answer, the doc_id, and the evidence all refer to the same incident and the same report date?
- **Question type correctness:** is the item labelled with the correct family (situational / lifeline / complex) and, for complex items, the correct category?

- **Unanswerable status validity:** is is_impossible set correctly? An item is unanswerable only when the record is genuinely silent; an explicit "zero / no impact / no threat" is answerable, and a generic-but-affirmative statement is not unanswerable.

Overall decision. After the six dimensions, the reviewer recorded keep, revise, or remove, and for revise supplied the corrected question, answer, or evidence. The two independent passes were then adjudicated by the specialist.

A.2 Evaluation Details

A.2.1 Reader Prompts and Output Schema for the Five Conditions

Every model is evaluated under five conditions that differ in the evidence packed into the prompt. All conditions share one strict one-line JSON output schema; the reader must return its answer, the field identifiers it used, and an explicit answerable flag. Every evidence block (gold, retrieved, temporal distractor, or full incident corpus) is labelled with its field identifier in the identical format [<box_uid>], so the gold block is never identifiable by a privileged label and citations resolve against the exact identifiers shown. Prompts are identical across all models:

```
System prompt (evidence conditions: gold evidence, retrieval, temporal distractor, full
     incident corpus).
You answer questions about U.S. wildfire ICS-209 incident reports. Base your answer ONLY on
     evidence provided (if any). If the question cannot be answered from the material, set
     answerable=false and leave answer empty. Do not invent facts, boxes, or dates.
Respond with STRICT JSON on one line and nothing else:
{"answer": "<text or empty>", "cited_boxes": ["<box_uid>", ...], "answerable": true|false}
System prompt (closed-book condition).
You answer questions about U.S. wildfire ICS-209 incident reports from your own knowledge
     only (no documents are provided). If you do not know, set answerable=false and leave
     answer empty.
Respond with STRICT JSON on one line and nothing else:
{"answer": "<text or empty>", "answerable": true|false}
User message. One shared template; conditions differ only in which evidence blocks are
     packed, each rendered as [<label>] then its text.
Evidence boxes (cite the box_uid(s) you use):
{rendered evidence blocks, each  [<box_uid>]  then the box text}

Question: {question}

# closed-book: user message is exactly  "Question: {question}"  (no evidence header)
# gold evidence: one block per gold box; EMPTY context when the item is unanswerable (true
     abstention)
# retrieval: the hybrid retriever's top-10 boxes for the incident, one block each
# temporal: gold boxes + the same boxes from adjacent report dates, then deterministically
     shuffled by a sha256("temporal:"+qa_uid) seed so the correct block is not first
# full-incident-corpus: the entire corpus (PDFs) of the question's incident (no retriever)
```

Output schema. `answer`: string (empty when abstaining); `cited_boxes`: list of box_uid strings (evidence conditions only); `answerable`: boolean. Responses are parsed into a response tag (answered / abstained / empty_answerable / malformed_json / refusal / gen_error); only an "answered" response on an answerable item is scored as an answer, and claims are decomposed downstream by the judge rather than emitted by the reader.

A.2.2 Metric Definitions

All metrics are computed by the aggregation code; means carry bootstrap 95% confidence intervals (1,000 resamples). Citations and evidence-group scores are resolved against the set of field identifiers actually exposed in the prompt, so a citation of a field that was not shown counts as a fabrication even if that field is in the gold set.

Answer correctness.

- Per-item score: each judge label maps incorrect/partial/correct to $0/0.5/1$; a judged item's score is the symmetric mean of the two judges, or the adjudicated label when a third pass exists: $\text{score} = \dfrac{judge_1 + judge_2}{2}$.

- Correctness: mean per-item score over all answerable items (an abstained or empty answer contributes 0).
- Quality | answered / strict: mean score over items actually answered; strict correctness = fraction where both judges independently said "correct".

Evidence citation.

- Citation attempt rate: fraction of answered items that returned at least one field identifier: $\mathbf{1}_{n_{\text{cited}} \geq 1}$.
- Annotated-field precision: returned identifiers that are in the gold set, over identifiers returned: $\dfrac{|\text{cited} \cap \text{gold}|}{n_{\text{cited}}}$.
- Annotated-field recall: coverage of the gold set that was actually exposed, best over sufficient evidence groups g: $\max_g \dfrac{|\text{cited} \cap (g \cap \text{exposed})|}{|g \cap \text{exposed}|}$.
- Annotated-field F1: harmonic mean of precision P and recall R: $\text{F1} = \dfrac{2\,P\,R}{P+R}$.
- Evidence-group success: 1 if some sufficient evidence group is both fully shown and fully cited: $\mathbf{1}_{\exists\, g\,:\, g \subseteq \text{exposed} \,\wedge\, g \subseteq \text{cited}}$.
- Joint answer and evidence correctness: 1 if the answer is graded correct and a complete sufficient group is cited (the strictest measure): $\mathbf{1}_{\text{correct} \,\wedge\, \text{group_covered}}$.
- Fabricated citation rate: returned identifiers that appear nowhere in the prompt, over identifiers returned: $\dfrac{|\text{cited} \notin \text{prompt}|}{n_{\text{cited}}}$.

Claim support.

- The answer is split into atomic claims, each labelled entailed / contradicted / neutral against the evidence; unsupported-claim rate $= 1-$ supported precision: supp. precision $= \dfrac{\text{entailed}}{\text{entailed} + \text{contradicted}}$, contradiction $= \dfrac{\text{contradicted}}{\text{all claims}}$.

Abstention.

- Confusion-matrix rates: the positive class is "should abstain"; a response counts as abstaining if answerable=false or the response is tagged abstained/refusal: $P = \dfrac{tp}{tp+fp}$, $R = \dfrac{tp}{tp+fn}$, $\text{F1} = \dfrac{2PR}{P+R}$.
- False-answer rate: unanswerable items that were answered, over the 11 unanswerable items: $\dfrac{\text{false_answer}}{11}$.
- Over-abstention rate: answerable items abstained on, over answerable items: $\dfrac{fp}{n_{\text{answerable}}}$.

Temporal robustness. An item is *temporal-eligible* if at least one of its gold fields takes a differing value at an adjacent reporting period, so that the *Gold with Temporal Distractors* condition presents a conflicting alternative rather than a duplicate of the gold box. Every temporal metric below is computed on this subset only. Over eligible items scorable in both conditions, the temporal robustness is answer correctness under *Gold Evidence* minus answer correctness under *Gold with Temporal Distractors*.

Retrieval (offline). Recall@k with gold fields absent from the extracted corpus removed from the denominator (coverage-aware); MRR and nDCG@k are standard; evidence-group success@k is 1 if the top-k contains a complete sufficient group. Recall@$k = \dfrac{|\text{top-}k \cap \text{gold}|}{|\text{gold}|}$.

A.2.3 Retrieval Recall Across Cut-offs

We compare three retrievers. The sparse retriever is Okapi BM25 with the standard parameters k_1 of 1.5 and b of 0.75, left untuned and with stopwords retained. The dense retriever is BAAI bge-small-en-v1.5 with the recommended query prefix, a maximum sequence length of 512 and cosine similarity over L2 normalized embeddings. The hybrid retriever fuses the two with reciprocal rank fusion at a constant of 60 over the top 50 candidates from each. Figure A.1 reports Field Recall@k for the three retrievers over answerable questions. Evidence-group success@k, which requires a complete sufficient group in the top-k rather than a single field, is strictly lower (0.342 at $k = 10$ for hybrid) because complex questions need several fields retrieved together. Because all retrieval hyper-parameters are fixed a priori and left untuned, this curve is a property of the benchmark rather than of a tuned system, and it quantifies how hard it is to locate the right field on the right day within a single incident.

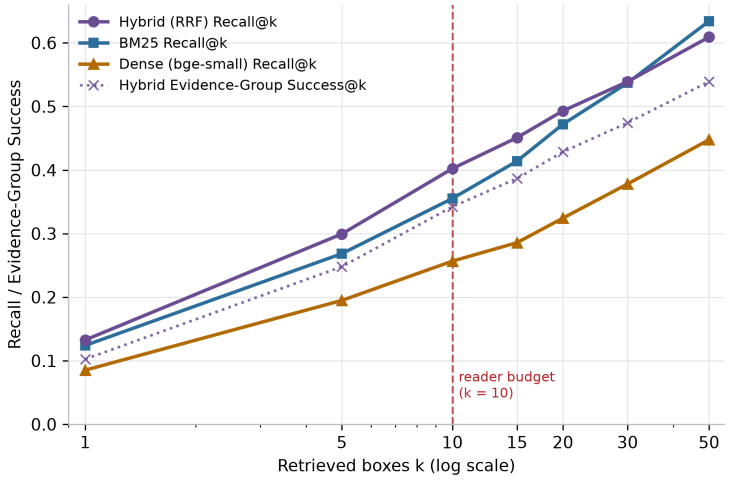

Figure A.1: Field Recall@k for BM25, dense and hybrid retrieval on a logarithmic k-axis, with the hybrid Evidence-Group Success@k and the reader budget ($k = 10$) marked.

Extending the analysis to the full corpus depth (with an unbounded RRF pool) confirms that recall climbs toward 1 only as k approaches the incident size (Figure A.2).

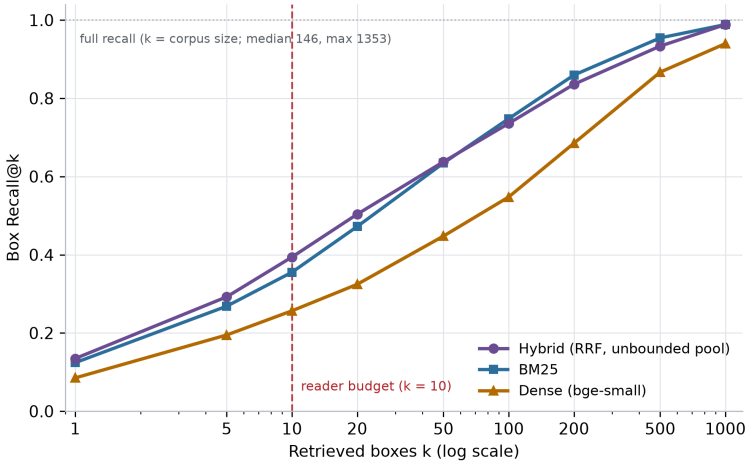

Figure A.2: Field Recall@k extended to $k = 1{,}000$. Recall saturates near the incident corpus size.

Table A3: Conditional answer quality and grading coverage. Each cell reports the number of answered questions graded (n), followed by conditional quality (Q), and strict correctness (S).

Model	Gold Evidence n / Q / S	Retrieval n / Q / S	Gold with Temporal Distractors n / Q / S	Full-Incident n / Q / S
Gemini-3.1-Pro-Preview	721 / 0.947 / 0.795	346 / 0.809 / 0.630	724 / 0.939 / 0.773	692 / 0.923 / 0.751
Gemini-3.7-Flash	701 / 0.974 / 0.865	321 / 0.846 / 0.707	705 / 0.977 / 0.867	685 / 0.939 / 0.785
Gemini-3.8-Flash	678 / 0.981 / 0.870	286 / 0.902 / 0.752	676 / 0.980 / 0.858	612 / 0.943 / 0.820
Gemini-3.5-Flash-Lite	700 / 0.951 / 0.770	348 / 0.805 / 0.629	702 / 0.936 / 0.766	607 / 0.836 / 0.616
GPT-OSS-120B	699 / 0.930 / 0.717	355 / 0.775 / 0.558	696 / 0.925 / 0.724	–
Llama4-Maverick-17B	735 / 0.915 / 0.615	433 / 0.756 / 0.499	735 / 0.909 / 0.657	741 / 0.772 / 0.429
Llama4-Scout-17B	718 / 0.903 / 0.618	403 / 0.736 / 0.496	721 / 0.904 / 0.619	718 / 0.697 / 0.366
Gemma3-27B-it	764 / 0.930 / 0.628	605 / 0.641 / 0.370	768 / 0.913 / 0.617	–
Gemma4-31B-it	725 / 0.970 / 0.764	408 / 0.798 / 0.564	730 / 0.964 / 0.740	710 / 0.860 / 0.594
GLM-5-2	688 / 0.983 / 0.874	364 / 0.827 / 0.654	707 / 0.986 / 0.853	630 / 0.958 / 0.782

A.2.4 JUDGE RUBRIC

Answer correctness is graded by two independent LLM judges, Claude Opus 4.8 and GPT-5.6 Sol, each blinded to the identity of the answering model and each given the identical rubric below:

```
You are a STRICT, NEUTRAL grader for a U.S. wildfire ICS-209 question-answering benchmark.
     For each item you are given a QUESTION, the GOLD_ANSWER (the reference correct answer),
      and a CANDIDATE_ANSWER produced by an anonymous system. Judge how well the candidate
     matches the gold answer in MEANING, not wording. Use exactly three labels:
  correct   = conveys all essential facts of the gold (right entities, numbers, dates, and
     polarity/negation); paraphrase, formatting, or extra correct detail is fine; nothing
     contradicts gold.
  partial   = the core is right but an essential fact is missing or minorly wrong (e.g. an
     incomplete list, one wrong secondary number), OR right core with a contradicting detail
     .
  incorrect = wrong, contradicts the gold, or does not answer the question.
For a list/multi-item gold, 'correct' needs substantially all items; 'partial' if only some.
      Grade ONLY against the gold answer and question; do not use outside knowledge. Output
     STRICT JSON and NOTHING else: a JSON array with one object per item, in the same order:
[{"id": "<id>", "verdict": "correct|partial|incorrect", "reason": "<=12 words"}]
```

A.3 ADDITIONAL EVALUATION RESULTS

A.3.1 ANSWER QUALITY AND EVALUATION COVERAGE

Table A3 separates answer coverage from quality. Conditional quality averages the final scores of answered questions, whereas strict correctness records the fraction for which both original judges independently assign a correct label. Neither measure includes unanswered questions. The headline correctness in Table 3 also assigns zero to nonanswers and therefore accounts for useful answer coverage.

A.3.2 EVIDENCE CITATION

Table A4 reports citation metrics over answered responses. Joint correctness in Table 3 uses all 789 answerable questions. The distinction matters because a model may cite evidence well on a small subset while leaving many questions unanswered. Annotated-field recall is restricted to gold identifiers exposed in the prompt. High citation recall can occur when only part of the required evidence is supplied. Evidence group success addresses this limitation by requiring a complete sufficient group to be both present and cited.

A.3.3 TEMPORAL ROBUSTNESS

An item is eligible for temporal evaluation when an adjacent report contains a differing value in at least one relevant gold field. Table A5 reports the paired correctness difference on eligible items

that are scorable in both the *Gold Evidence* and *Gold with Temporal Distractors* conditions. Positive values indicate lower correctness with distractors, and negative values indicate an increase.

Table A4: Evidence citation performance on answered responses. Each cell reports annotated-field F1, evidence group success (EGS), and fabricated citation rate (F). Higher F1 and EGS and lower F indicate better performance.

Model	Gold Evidence F1 / EGS / F	Retrieval F1 / EGS / F	Gold with Temporal Distractors F1 / EGS / F	Full-Incident F1 / EGS / F
Gemini-3.1-Pro-Preview	0.973 / 0.910 / 0.000	0.782 / 0.627 / 0.000	0.967 / 0.919 / 0.001	0.745 / 0.702 / 0.053
Gemini-3.7-Flash	0.982 / 0.943 / 0.004	0.807 / 0.676 / 0.000	0.977 / 0.963 / 0.007	0.740 / 0.739 / 0.011
Gemini-3.8-Flash	0.987 / 0.950 / 0.000	0.847 / 0.731 / 0.000	0.979 / 0.962 / 0.004	0.769 / 0.763 / 0.008
Gemini-3.5-Flash-Lite	0.976 / 0.927 / 0.001	0.755 / 0.609 / 0.014	0.960 / 0.912 / 0.003	0.573 / 0.476 / 0.269
GPT-OSS-120B	0.912 / 0.863 / 0.076	0.575 / 0.465 / 0.167	0.865 / 0.806 / 0.096	–
Llama4-Maverick-17B	0.966 / 0.903 / 0.004	0.532 / 0.443 / 0.163	0.888 / 0.803 / 0.033	0.425 / 0.389 / 0.365
Llama4-Scout-17B	0.957 / 0.875 / 0.000	0.671 / 0.526 / 0.017	0.920 / 0.836 / 0.001	0.298 / 0.258 / 0.008
Gemma3-27B-it	0.952 / 0.856 / 0.002	0.477 / 0.374 / 0.049	0.920 / 0.828 / 0.001	–
Gemma4-31B-it	0.960 / 0.873 / 0.000	0.702 / 0.534 / 0.001	0.940 / 0.838 / 0.000	0.751 / 0.623 / 0.008
GLM-5-2	0.989 / 0.968 / 0.000	0.727 / 0.604 / 0.000	0.976 / 0.956 / 0.000	0.721 / 0.748 / 0.016

Table A5: Paired correctness changes under temporal distraction. Difference is answer correctness under *Gold Evidence* minus answer correctness under *Gold with Temporal Distractors*. Counts identify the scorable eligible pairs used for each model.

Model	Paired correctness difference	Scorable eligible items
Gemini-3.1-Pro-Preview	0.012	523
Gemini-3.7-Flash	−0.001	511
Gemini-3.8-Flash	0.002	490
Gemini-3.5-Flash-Lite	0.015	504
GPT-OSS-120B	0.000	472
Llama4-Maverick-17B	0.009	530
Llama4-Scout-17B	−0.001	517
Gemma3-27B-it	0.020	561
Gemma4-31B-it	0.009	525
GLM-5-2	−0.004	499

A.3.4 Claim Support on Lifeline and Complex Questions

Claim support is evaluated on answered lifeline and complex questions. Table A6 shows supported claim precision, contradiction rate, and the number of evaluated answers. *Gold Evidence* produces higher supported claim precision than *Retrieval* for every model.

A.3.5 Abstention

Table A7 reports abstention F1, false answer rate, and over abstention. Under *Retrieval*, abstention F1 ranges from 0.042 to 0.079, alongside substantial abstention on answerable questions.

A.3.6 Performance by Question Type

Table A8 and Table A9 split the answer correctness and joint correctness by question type. The *Closed-Book* condition is omitted because answer correctness is at most 0.07 for every type and model. Type means average each question's score over the models in a condition.

Complex questions are answered as accurately as situational questions when the required evidence is available. Complex questions fall below situational questions only under *Retrieval*. This pattern is consistent with the need to retrieve several fields or reporting periods together for a complex answer.

Table A6: Claim support on answered lifeline and complex questions. Each cell reports supported claim precision (P), and contradiction rate (K), followed by the number of evaluated answers (n). Both scores are means of per answer fractions.

Model	Gold Evidence	Retrieval	Gold with Temporal Distractors	Full-Incident
	P / K (n)	P / K (n)	P / K (n)	P / K (n)
Gemini-3.1-Pro-Preview	0.799 / 0.163 (422)	0.694 / 0.205 (172)	0.760 / 0.183 (422)	0.657 / 0.180 (395)
Gemini-3.7-Flash	0.693 / 0.199 (408)	0.542 / 0.234 (160)	0.580 / 0.213 (409)	0.545 / 0.217 (389)
Gemini-3.8-Flash	0.719 / 0.202 (392)	0.630 / 0.197 (136)	0.626 / 0.212 (390)	0.584 / 0.224 (335)
Gemini-3.5-Flash-Lite	0.819 / 0.149 (402)	0.644 / 0.237 (175)	0.771 / 0.172 (402)	0.626 / 0.247 (337)
GPT-OSS-120B	0.744 / 0.166 (403)	0.481 / 0.275 (188)	0.675 / 0.181 (409)	–
Llama4-Maverick-17B	0.754 / 0.193 (429)	0.587 / 0.275 (225)	0.704 / 0.228 (426)	0.585 / 0.261 (431)
Llama4-Scout-17B	0.736 / 0.202 (418)	0.613 / 0.264 (215)	0.744 / 0.199 (419)	0.476 / 0.312 (412)
Gemma3-27B-it	0.790 / 0.171 (448)	0.533 / 0.262 (347)	0.770 / 0.188 (450)	–
Gemma4-31B-it	0.807 / 0.148 (423)	0.560 / 0.236 (222)	0.783 / 0.148 (426)	0.622 / 0.198 (406)
GLM-5-2	0.780 / 0.173 (395)	0.616 / 0.220 (185)	0.731 / 0.192 (404)	0.580 / 0.239 (336)

Table A7: Abstention performance across evidence conditions. Each cell reports abstention F1, false answer rate (FA), and over-abstention (OA). Higher F1 and lower FA and OA indicate better performance.

Model	Gold Evidence	Retrieval	Gold with Temporal Distractors	Full-Incident
	F1 / FA / OA	F1 / FA / OA	F1 / FA / OA	F1 / FA / OA
Gemini-3.1-Pro-Preview	0.250 / 0.000 / 0.084	0.048 / 0.000 / 0.559	0.259 / 0.000 / 0.080	0.204 / 0.091 / 0.098
Gemini-3.7-Flash	0.200 / 0.000 / 0.112	0.045 / 0.000 / 0.593	0.209 / 0.000 / 0.105	0.167 / 0.091 / 0.126
Gemini-3.8-Flash	0.172 / 0.000 / 0.134	0.042 / 0.000 / 0.632	0.169 / 0.000 / 0.137	0.117 / 0.091 / 0.190
Gemini-3.5-Flash-Lite	0.198 / 0.000 / 0.113	0.048 / 0.000 / 0.559	0.202 / 0.000 / 0.110	0.099 / 0.091 / 0.231
GPT-OSS-120B	0.242 / 0.000 / 0.087	0.048 / 0.091 / 0.502	0.250 / 0.000 / 0.084	–
Llama4-Maverick-17B	0.289 / 0.000 / 0.068	0.048 / 0.182 / 0.446	0.289 / 0.000 / 0.068	0.194 / 0.455 / 0.057
Llama4-Scout-17B	0.237 / 0.000 / 0.090	0.044 / 0.182 / 0.488	0.244 / 0.000 / 0.086	0.212 / 0.182 / 0.082
Gemma3-27B-it	0.489 / 0.000 / 0.029	0.079 / 0.273 / 0.232	0.537 / 0.000 / 0.024	–
Gemma4-31B-it	0.259 / 0.000 / 0.080	0.050 / 0.091 / 0.482	0.272 / 0.000 / 0.075	0.210 / 0.091 / 0.094
GLM-5-2	0.182 / 0.000 / 0.126	0.050 / 0.000 / 0.534	0.220 / 0.000 / 0.099	0.167 / 0.091 / 0.126

Lifeline questions have the lowest answer correctness in every evidence condition. Most of this gap comes from abstention rather than wrong answers. Under *Gold Evidence*, models abstain on 13.5% of lifeline questions, compared with 6.6% of situational and 7.4% of complex questions, while answers scored 0 remain below 1% for all three types.

Complex questions are harder to ground in evidence than to answer when the evidence must be located. Their mean joint correctness is lower than that of situational questions in every evidence condition, and the difference is clear under *Retrieval* and *Full-Incident* but small when the gold evidence is supplied. Adjacent reporting periods do not lower complex joint correctness, whereas situational and lifeline joint correctness decrease slightly.

Table A8: Answer correctness by question type. Each cell reports situational / lifeline / complex answer correctness (S / L / C). The last row averages each question's score over the models in each condition.

Model	Gold Evidence S / L / C	Retrieval S / L / C	Gold with Temporal Distractors S / L / C	Full-Incident S / L / C
Gemini-3.1-Pro-Preview	0.883 / 0.823 / 0.928	0.461 / 0.249 / 0.362	0.891 / 0.808 / 0.920	0.870 / 0.740 / 0.833
Gemini-3.7-Flash	0.892 / 0.808 / 0.938	0.461 / 0.243 / 0.315	0.909 / 0.819 / 0.920	0.883 / 0.728 / 0.866
Gemini-3.8-Flash	0.878 / 0.785 / 0.899	0.441 / 0.243 / 0.264	0.872 / 0.793 / 0.877	0.817 / 0.634 / 0.764
Gemini-3.5-Flash-Lite	0.889 / 0.801 / 0.844	0.470 / 0.255 / 0.326	0.878 / 0.784 / 0.844	0.736 / 0.505 / 0.761
GPT-OSS-120B	0.878 / 0.799 / 0.757	0.433 / 0.279 / 0.319	0.848 / 0.804 / 0.772	–
Llama4-Maverick-17B	0.898 / 0.785 / 0.906	0.520 / 0.299 / 0.449	0.895 / 0.772 / 0.917	0.778 / 0.644 / 0.797
Llama4-Scout-17B	0.852 / 0.781 / 0.851	0.466 / 0.273 / 0.413	0.853 / 0.767 / 0.902	0.719 / 0.517 / 0.721
Gemma3-27B-it	0.925 / 0.863 / 0.935	0.580 / 0.402 / 0.504	0.911 / 0.847 / 0.935	–
Gemma4-31B-it	0.914 / 0.838 / 0.964	0.497 / 0.314 / 0.453	0.916 / 0.831 / 0.982	0.841 / 0.660 / 0.891
GLM-5-2	0.900 / 0.799 / 0.899	0.492 / 0.272 / 0.388	0.930 / 0.826 / 0.913	0.903 / 0.651 / 0.717
Mean over models	0.891 / 0.808 / 0.892	0.482 / 0.283 / 0.379	0.890 / 0.805 / 0.898	0.818 / 0.635 / 0.794

Table A9: Joint answer and evidence correctness by question type. Each cell reports situational / lifeline / complex joint correctness (S / L / C). The last row averages each question's score over the models in each condition.

Model	Gold Evidence S / L / C	Retrieval S / L / C	Gold with Temporal Distractors S / L / C	Full-Incident S / L / C
Gemini-3.1-Pro-Preview	0.812 / 0.707 / 0.819	0.372 / 0.166 / 0.174	0.812 / 0.689 / 0.819	0.691 / 0.465 / 0.486
Gemini-3.7-Flash	0.853 / 0.722 / 0.848	0.378 / 0.172 / 0.196	0.884 / 0.758 / 0.855	0.728 / 0.492 / 0.558
Gemini-3.8-Flash	0.853 / 0.731 / 0.819	0.378 / 0.172 / 0.196	0.844 / 0.740 / 0.826	0.694 / 0.423 / 0.507
Gemini-3.5-Flash-Lite	0.819 / 0.695 / 0.761	0.381 / 0.157 / 0.159	0.772 / 0.668 / 0.768	0.459 / 0.190 / 0.239
GPT-OSS-120B	0.775 / 0.628 / 0.601	0.294 / 0.130 / 0.101	0.716 / 0.601 / 0.543	–
Llama4-Maverick-17B	0.791 / 0.637 / 0.746	0.319 / 0.139 / 0.159	0.728 / 0.550 / 0.790	0.381 / 0.242 / 0.246
Llama4-Scout-17B	0.728 / 0.634 / 0.703	0.328 / 0.154 / 0.145	0.719 / 0.583 / 0.703	0.263 / 0.088 / 0.196
Gemma3-27B-it	0.831 / 0.701 / 0.754	0.394 / 0.172 / 0.188	0.784 / 0.644 / 0.790	–
Gemma4-31B-it	0.816 / 0.698 / 0.841	0.391 / 0.187 / 0.167	0.794 / 0.671 / 0.804	0.634 / 0.347 / 0.478
GLM-5-2	0.869 / 0.773 / 0.819	0.394 / 0.184 / 0.188	0.884 / 0.789 / 0.870	0.762 / 0.417 / 0.478
Mean over models	0.815 / 0.693 / 0.771	0.363 / 0.163 / 0.167	0.794 / 0.669 / 0.777	0.577 / 0.333 / 0.399