NEOQA: Evidence-based Question Answering with Generated News Events

Max Glockner[1*], **Xiang Jiang**[2], **Leonardo F. R. Ribeiro**[2], **Iryna Gurevych**[1], **Markus Dreyer**[2]

[1]UKP Lab, ATHENE National Research Center for Applied Cybersecurity, TU Darmstadt
[2]Amazon AGI

{maxg216}@gmail.com {jxiang,leonribe,mddreyer}@amazon.com

Abstract

Evaluating Retrieval-Augmented Generation (RAG) in large language models (LLMs) is challenging because benchmarks can quickly become stale. Questions initially requiring retrieval may become answerable from pretraining knowledge as newer models incorporate more recent information during pretraining, making it difficult to distinguish evidence-based reasoning from recall. We introduce NEOQA (News Events for Out-of-training Question Answering), a benchmark designed to address this issue. To construct NEOQA, we generated timelines and knowledge bases of fictional news events and entities along with news articles and Q&A pairs to prevent LLMs from leveraging pretraining knowledge, ensuring that no prior evidence exists in their training data. We propose our dataset as a new platform for evaluating evidence-based question answering, as it requires LLMs to generate responses exclusively from retrieved evidence and only when sufficient evidence is available. NEOQA enables controlled evaluation across various evidence scenarios, including cases with missing or misleading details. Our findings indicate that LLMs struggle to distinguish subtle mismatches between questions and evidence, and suffer from short-cut reasoning when key information required to answer a question is missing from the evidence, underscoring key limitations in evidence-based reasoning.[1]

1 Introduction

Retrieval-Augmented Generation (RAG) equips LLMs with external information to complement their parametric knowledge (Lewis et al., 2020; Yu et al., 2025) and enables them to answer questions that involve information beyond their pretraining data, such as recent events or rare entities. For trustworthy applications, the ability to reason over multiple evidence documents is critical to producing verifiable answers grounded in these documents (Liu et al., 2023; Yue et al., 2023; Li et al., 2024b). LLMs must not only be able to answer a question correctly when the evidence is sufficient, but also be able to deflect from answering if the question cannot be answered given the evidence (Cao, 2024). However, benchmarks for evidence-based reasoning on real-world data lose value over time as LLMs increasingly can rely on updated parametric knowledge from pretraining rather than external information (Figure 1, *right*), as analyzed empirically in Section 2. Constructing datasets with recent data (Chen et al., 2024; Tang and Yang, 2024; Karpinska et al., 2024) only postpones the issue until LLMs are retrained, while frequent dataset updates (Vu et al., 2024; Kasai et al., 2024) mitigate it but make consistent progress tracking difficult.

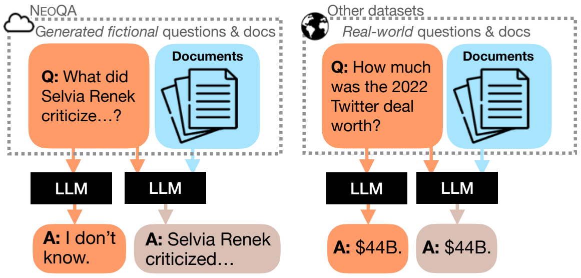

NEOQA
Generated fictional questions & docs
Q: What did Selvia Renek criticize…?
Documents
LLM
LLM
A: I don't know.
A: Selvia Renek criticized…
Other datasets
Real-world questions & docs
Q: How much was the 2022 Twitter deal worth?
Documents
LLM
LLM
A: $44B.
A: $44B.

Figure 1: **Left:** NEOQA features LLM-generated questions and documents about events from a fictional timeline, ensuring that LLMs can only answer by reasoning over the documents. **Right:** Real-world RAG datasets become ineffective for newer LLMs that have internalized knowledge of recent events, rendering the provided evidence documents redundant.

To address these challenges, we introduce NEOQA, a fully LLM-generated dataset of fictional events with associated news articles, as well as question-answer pairs. Organized into *timelines*, each with ten sequential *events* (Figure 2), NEOQA mimics how events unfold over time in reality. All events and named entities are fictional to avoid

[*] Work was done while MG was an intern at Amazon AGI.
[1] https://github.com/amazon-science/neoqa

Findings of the Association for Computational Linguistics: ACL 2025, pages 11842–11926
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

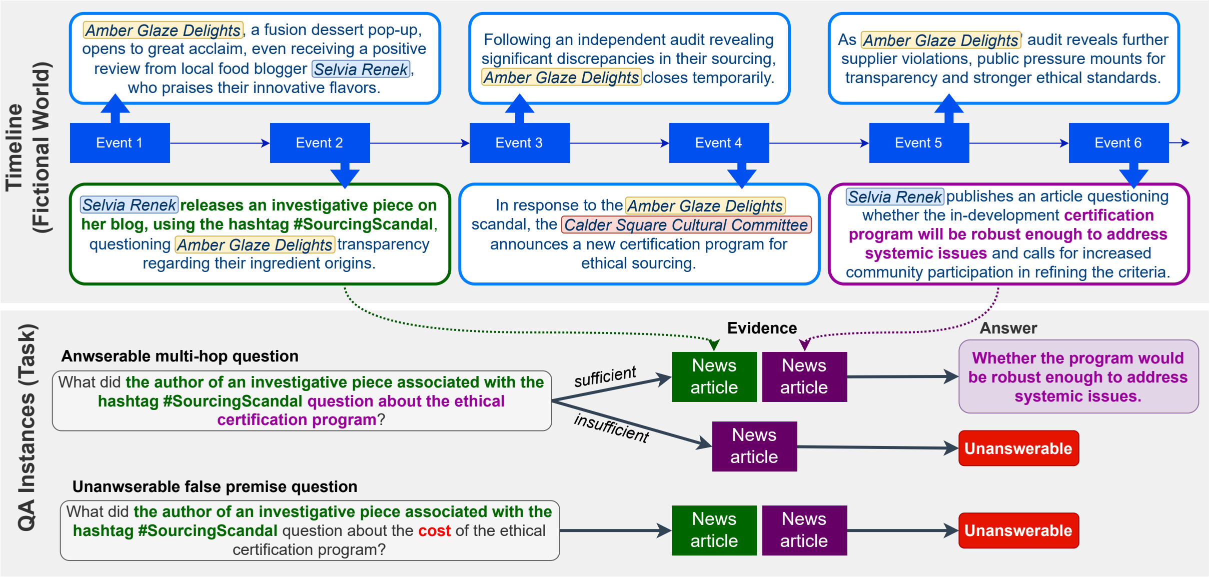

Timeline (Fictional World)
Amber Glaze Delights, a fusion dessert pop-up, opens to great acclaim, even receiving a positive review from local food blogger Selvia Renek, who praises their innovative flavors.
Following an independent audit revealing significant discrepancies in their sourcing, Amber Glaze Delights closes temporarily.
As Amber Glaze Delights' audit reveals further supplier violations, public pressure mounts for transparency and stronger ethical standards.
Event 1
Event 2
Event 3
Event 4
Event 5
Event 6
Selvia Renek releases an investigative piece on her blog, using the hashtag #SourcingScandal, questioning Amber Glaze Delights transparency regarding their ingredient origins.
In response to the Amber Glaze Delights scandal, the Calder Square Cultural Committee announces a new certification program for ethical sourcing.
Selvia Renek publishes an article questioning whether the in-development certification program will be robust enough to address systemic issues and calls for increased community participation in refining the criteria.
QA Instances (Task)
Evidence
Answer
Anwserable multi-hop question
What did the author of an investigative piece associated with the hashtag #SourcingScandal question about the ethical certification program?
sufficient
insufficient
News article
News article
News article
Whether the program would be robust enough to address systemic issues.
Unanswerable
Unanwserable false premise question
What did the author of an investigative piece associated with the hashtag #SourcingScandal question about the cost of the ethical certification program?
News article
News article
Unanswerable

Figure 2: An extract of a timeline from NEOQA with six out of ten events (summarized for visualization) with highlighted fictional named entities. Answering a multi-hop question requires combining information from two events. The model should deflect when only partial (*insufficient*) information is available or when subtle permutations make the question unanswerable (e.g., false premise questions).

interference from LLMs with updated parametric knowledge (Figure 1, *left*). Each event includes resolved named entities, and a corresponding knowledge base (KB) entry is created and continuously updated as the events progress. The events adhere to real-world physical laws and common sense, allowing models to leverage their commonsense reasoning (Choi, 2022). For each event, news articles and multiple-choice questions are independently generated and grounded in identifiable atomic information. This allows to pair a question with any set of news articles and clearly distinguish between sufficient or insufficient evidence, and unrelated documents. *News articles* serve as evidence and focus on different information of a single event, *questions* require reasoning over information from up to two previous events. For example, answering the question in Figure 2 requires combining the red and purple facts. Any set of news articles with both facts provides sufficient evidence, while any with fewer is insufficient. This allows us to test models under different evidence conditions, requiring them to answer correctly when sufficient evidence is available and to deflect when it is not. Overall, by grounding NEOQA's news articles—used as inference-time evidence—and questions in fictional timelines that are independent from real-world news cycles, NEOQA remains a reliable benchmark for evaluating future LLMs, free from the risk of pretraining data contamination or knowledge conflicts about real-world events. Following the recommendations of Jacovi et al. (2023), we release NEOQA under a no-derivatives license (CC-BY-ND-4.0) and with public key encryption.

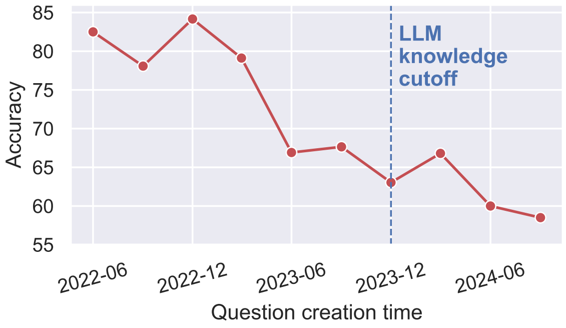


Figure 3: GPT-4 Turbo accuracy on RealTimeQA questions (no RAG evidence provided). It answers older questions more accurately from memory, suggesting that older RAG datasets can be solved without RAG.

Because parametric knowledge cannot replace external evidence in NEOQA, a question can only be answered correctly when sufficient external evidence is provided. When evidence is insufficient, the model can only *guess* the correct answer using shortcut reasoning (Jiang and Bansal, 2019; Chen and Durrett, 2019; Trivedi et al., 2022), undermining their trustworthiness. Determining whether a model is using shortcut reasoning to guess an answer or genuinely completing it with its learned (parametric) knowledge is difficult in real-world datasets, where models can often justifiably fill in knowledge gaps. However, this is not the case in NeoQA, where answers require sufficient external evidence. This setup makes it possible to identify and penalize shortcut reasoning during evaluation. For example, in HotpotQA (Yang et al., 2018), answering "Shenley Hall is a house in a parish how far from central London?" requires identifying the house's village and its distance from London. If only the latter is provided, it is unclear whether the model reasons correctly or uses shortcuts. NeoQA enables controlled experiments on evidence-based reasoning that account for shortcut reasoning, requiring models to compare questions with evidence and answer only when justified, deflecting otherwise. It includes answerable and unanswerable questions (with unverifiable or incorrect assumptions (Kim et al., 2021; Hu et al., 2023)), combined with evidence that is sufficient, insufficient, and/or distracting. Our experiments show that models struggle to distinguish sufficient from insufficient evidence, frequently relying on shortcuts and failing to detect subtle mismatches. In summary, our contributions are:

1. A **novel methodology** for automatically generating an evidence-based question-answering dataset grounded in fictional timelines.
2. The **NeoQA dataset** with diverse question types and evidence configurations for evaluating evidence-based reasoning.
3. **Controlled experiments** reveal the challenges posed by shortcut reasoning.

2 Parametric Knowledge Interference

Data contamination in LLM pretraining, where test data overlaps with training data, has compromised several benchmarks (Magar and Schwartz, 2022; Jacovi et al., 2023; Elazar et al., 2024; Sainz et al., 2024). We test whether RAG benchmarks are similarly affected by events overlapping with the pretraining data. If LLMs acquire relevant knowledge during pretraining, such benchmarks lose their purpose. To quantify this, we evaluate GPT-4 Turbo (reported knowledge cutoff: December 2023[2]) on RealTimeQA (Kasai et al., 2024), a dataset of weekly multiple-choice news quizzes (June 2022–Jan 2024). We extend the dataset through September 2024 (see Appendix A) and test the model's accuracy using only its parametric knowledge, without external evidence. Figure 3 shows higher accuracy on older questions, indicating that the LLM acquired much of the relevant information during pretraining. Performance drops sharply around March 2023, several months before the reported knowledge cutoff. We hypothesize that this discrepancy arises because reported knowledge cutoffs are conservative estimates, while the effective cutoff for different sources may be earlier (Cheng et al., 2024). Performance on unseen news remains above chance (25% for selecting from four options), likely due to common-sense reasoning, which helps eliminate distractors.

3 Task Definition

We introduce NeoQA, a QA dataset agnostic to parametric knowledge by focusing on fictional events and named entities that do not exist in the real world. The task is formulated as multiple-choice, where the model receives a question, a preselected set of news articles as evidence, and a set of seven candidate answers (a correct answer, a deflection option if unanswerable, and five distractor answers). The model must assess whether the evidence is sufficient to answer the question, select the correct answer if possible, and deflect if the evidence is insufficient, or if the question's assumptions are unverifiable or incorrect. We always include an explicit "unanswerable" option, which has been shown to help models deflect when the answer is unknown (Slobodkin et al., 2023). We do not include document retrieval in our task formulation, but instead control the preselected evidence to simulate realistic retrieval conditions with sufficient, insufficient, or irrelevant information. This allows for a fine-grained evaluation of how models reason over imperfect evidence, as real-world retrieval often includes noise or missing details.

The multiple-choice approach addresses two key challenges in LLM-generated answers: (1) avoiding the need for an expensive judge LLM, and (2) reducing false negatives caused by ambiguous questions, which can lead to multiple different but valid interpretations based on different evidence (Min et al., 2020; Glockner et al., 2024). While we instructed the LLM to avoid such question ambiguity during dataset creation, this cannot be guaranteed, and exhaustive annotation across numerous evidence articles is impractical. Similar to challenge datasets (McCoy et al., 2019; Schuster et al., 2019; Gardner et al., 2020), we do not fine-tune

[2] https://platform.openai.com/docs/models

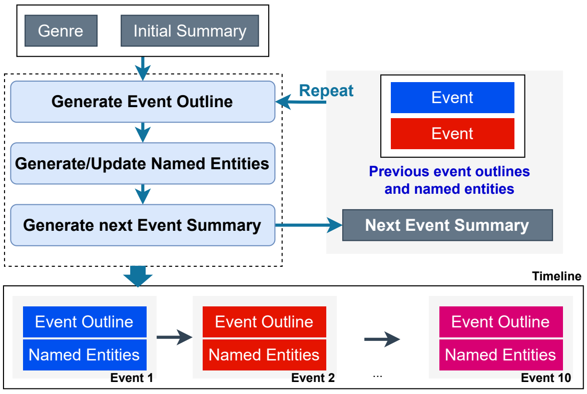


Figure 4: Events are generated sequentially based on a summary sentence and the previously generated events.

models on NeoQA to prevent them from reverse-engineering its specific question generation strategies. Instead, we define the task as zero-shot. However, we provide a separate development set for prompt selection to avoid overfitting to the test set.

4 NeoQA Construction

Our goal is to create a stable QA dataset based on fictional event timelines where LLMs with updated parametric knowledge have no unfair advantages. Creating NeoQA involves three steps: (1) generating independent fictional timelines of ten events, (2) writing news articles about the events, and (3) creating questions about them. Questions and news articles are independently generated and linked to specific information sentences in the events, allowing the automatic generation of instances with questions with sufficient or insufficient evidence, and with distracting documents. We use GPT-4o to generate all parts of NeoQA.

4.1 Timeline Generation

The fictional world of NeoQA uses independent timelines that mirror real-world event progression (Shahaf and Guestrin, 2010; Pratapa et al., 2023), following sequential events in narrative literature (Keith et al., 2023). Unlike tree-like structures with divergent subplots (Liu et al., 2017), this approach conditions each event on all previous events in the same storyline, reducing the risk of introducing inconsistencies. Each event includes a *date*, an *outline* describing the plot, and a KB of fictional *named entities*, which is updated after each event. Each outline contains 20-30 single-sentence *outline items*, each providing distinct details about the event. Examples are provided in Appendix B.

The timeline generation adapts steps from story generation methods (Yang et al., 2023; Lee et al., 2025; Zhu et al., 2023), but differs as it avoids classic narrative templates with protagonists and antagonists. Specifically, it begins with an LLM-generated seed summary and news genre, followed by three core steps using multiple prompts: (1) creating, checking, and refining the outline, (2) generating and updating fictional named entities, and (3) producing a new seed summary for the next event. We employ heuristics to detect and correct errors by critiquing the LLM outputs (Gou et al., 2024). Events are generated sequentially, conditioned on prior events, named entities, and the latest summary sentence (Figure 4). To help the LLM maintain consistency with named entities, we provide all past event outlines with resolved entities and updated knowledge base entries, mirroring neural representation approaches (Clark et al., 2018).

The boundary between a "realistic" fictional world and reality is blurred due to inevitably overlapping concepts. For example, common sense (e.g., what "rain" is) and physical laws must still apply in the fictional world, while unrealistic elements, like dragons, must not exist. We define two practical criteria for fictional worlds to minimize the impact of updated parametric knowledge from real-world data while preserving common sense and physical laws: a) fictional named entities and b) mutually exclusive sampling for seed summaries of subsequent events. First, NeoQA distinguishes seven types of named entities from Ling and Weld (2021) with one extra type for "miscellaneous". We compare each named entity against Wikipedia to ensure it doesn't overlap with well-known real-world entities. However, this alone is insufficient, as the LLM might generate timelines based on its parametric knowledge, aligning with real-world events but simply replacing the named entities. To prevent this, the LLM generates multiple mutually exclusive summaries for subsequent events, from which one is randomly selected. This way, the timeline follows irreversible paths and prevents the LLM from generating summaries that closely resemble real-world events from its parametric knowledge. For example, a possible (not chosen) seed summary for the second event in Figure 2 was, that after the initial accusations in the first event, Amber Glaze Delights suspended operations for an internal audit.

News articles about the events in each timeline serve as evidence documents in NeoQA. The LLM generates news articles with four profiles ("progressive" "conservative", "objective", "sensational") to

Type	Example Text
Evidence 1	Selvia Renek released an investigative piece on her blog questioning Amber Glaze Delights' transparency, which gained traction on social media under the hashtag #SourcingScandal.
Evidence 2	Selvia Renek published an article questioning whether the certification program would be **robust enough** to address systemic issues and called for increased community participation in refining the criteria.
Multi-Hop Question (✓)	What did the author of an investigative piece associated with the hashtag #SourcingScandal question about the ethical certification program?
False Premise (×)	What did the author of an investigative piece associated with the hashtag #SourcingScandal question **about the cost** of the ethical certification program?
Uncertain Specificity (×)	What did the author of an investigative piece associated with the hashtag #SourcingScandal question about the certification program's **ability to address exploitative labor practices**?

Table 1: An answerable multi-hop question (parent question) and unanswerable questions. The false premise question contradicts the evidence (**red**). The uncertain specificity question introduces unverifiable details (**blue**).

simulate how organizations emphasize different aspects in their reporting (Fan et al., 2019). For each profile, the model (1) selects three subsets of multiple outline items from each individual event outline, (2) drafts an article for each selection, and (3) verifies that all information from these outline items is included in the final article. The generation process is detailed in Appendix C.

4.2 Question and Answer Generation

NeoQA includes four question types. These questions require distinct evidence-based reasoning skills. The models must provide accurate answers when possible and detect when the answer cannot be determined due to unverifiable, contradictory, or missing evidence. (1) **Time-span** questions involve temporal reasoning to calculate the duration between outline items across up to two events (Example: *"How many days passed between the announcement of the public forum by the Calder Square Cultural Committee and the day the forum was held?"*). Both can be used to form answerable instances (with sufficient evidence) or unanswerable (with insufficient evidence). (2) **Multi-hop** questions use a fictional named entity as a bridge entity (Yang et al., 2018; Tang and Yang, 2024) to link information from two sentences (see example in Figure 2). From the multi-hop questions, we create two types of questions that are always unanswerable, regardless of the evidence: (3) **False premise** questions have incorrect assumptions that directly contradict the evidence. This differs from false premise questions in other works (Yang et al., 2024b) where the assumptions contradict general world knowledge. (4) **Uncertain specificity** questions ask for details that are too specific to be answered by the available evidence in the fictional timeline. Examples are shown in Table 1.

		Overall	Per Timeline
World	**Timelines**	15	1
	Events	150	10
	Outline Sentences	3,174	211.6
	Named Entities	393	26.2
Task	**Multi-hop**	839	55.9
	Time-span	678	45.2
	False premise	2,879	191.9
	Uncertain specificity	2,952	196.8
	News articles	1,800	150.0

Table 2: Summary statistics of elements in NeoQA.

To generate answerable questions, the LLM selects two outline items from up to two distinct events. Based on these outline items it then generates the question and correct answer, as well as five distractors, framing the question as if asked after the most recent event. We instruct the LLM to ensure (1) no other outline item can answer the question, and (2) both selected outline items are essential for a definite answer, with additional outline items added if needed. We provide all past event outlines as context to help the LLM avoid drafting ambiguous questions that could be answered differently using other information. For each multi-hop question, we instruct the LLM to generate multiple false premise and uncertain specificity questions using the same answer options, adding subtle contradictions or unverifiable details. These unanswerable questions share the same answer options as the original multi-hop question. Knowing what information is needed to answer a question and which news article contains it, allows us to combine questions with news articles to create scenarios where evidence is sufficient for an answer or insufficient, requiring the model to deflect. See Appendix C.2 for details on the generation process, and Figure 10 for a complete instance with question, answer options and news articles as evidence.

	Answerable	Unanswerable
Time-span	532	1,063
Multi-hop	625	1,239
False premise	–	1,250
Uncertain specificity	–	1,250
All instances	**1,157**	**4,802**

Table 3: Instances used for benchmarking experiments.

4.3 Quality Filtering and Assessment

To automatically link questions with news articles and determine their answerability, two key requirements must be met:

- **Requirement 1:** The selected outline items for creating each answerable question must be both fully sufficient and necessary to answer the question with certainty.
- **Requirement 2:** News articles must convey all factual information from the selected outline items and exclude any information from the non-selected outline items.

Requirement 1 can be violated when questions depend on information beyond the selected outline items, as the model had access to all events during generation. To mitigate this, we remove all 1,122 questions (42.5%) that the LLM, which generated the question, cannot answer correctly itself using only the selected outline items as evidence (see Appendix D.1). For Requirement 2, we use a pretrained T5 NLI model (Honovich et al., 2022) to verify that selected sentences are entailed by the news article, while unselected ones are not. The model agreed with the assumed entailment labels in 98.1% of selected sentences and 92.2% of unselected sentences. 7.3% of the remaining unselected sentences did not receive any label instead of a label disagreement (see Appendix D.2). Lastly, we conduct human annotation on 350 instances to verify the correctness of the reference answer. Majority voting from three annotators agreed with our labels 94% of the time. Fleiss's kappa of 0.516 indicates moderate agreement, underscoring the task's difficulty even for humans (see Appendix D.3 for details). Table 2 summarizes the statistics of NeoQA dataset after quality filtering. We use three timelines as the development set for prompt selection, and the rest twelve timelines as the test set.

5 Main Experiments

In our main experiments, we form multiple-choice instances by combining each question with all news articles available up to the question date. For multi-hop and time-span questions, we create one instance with complete evidence—where the model has access to all relevant information—and several instances with systematically reduced evidence. These *insufficient-evidence* instances are generated by omitting specific news articles, each corresponding to a different outline sentence necessary to answer the question. When the evidence is complete, the model is expected to select the correct answer; when it is insufficient, the correct response is to choose the deflection option. Following Schuster et al. (2021), this approach presents the same question with varying evidence, leading to different correct answers and requiring the model to use the evidence to perform well. Additionally, we include 2,500 instances with false premise and uncertain specificity questions, each paired with the same full set of news articles as the answerable instances. Details on instance creation are provided in Appendix E.2, with overall statistics in Table 3. Instances can include up to 120 documents, with a total of 1,349 to 45,484 tokens across both the question and evidence documents. This requires LLMs with sufficiently large context windows. We evaluate several open-source consumer-sized LLMs with up to 32B parameters: Qwen2.5 (Yang et al., 2024a) (7B, 14B, 32B), Phi3 (Abdin et al., 2024) (mini, small, medium), and Phi3.5 MoE. All these models support context sizes of at least 128k tokens. Prompts were selected per LLM using ADTScore (see below) on the development set (see Appendix F.1). Following Levy et al. (2024), we also test Chain-of-Thought (CoT) prompting (Wei et al., 2022) with the elicitation string (Zhou et al., 2023) to evaluate its effect on LLM performance. As our primary metric, we introduce ADTScore (Answer Deflection Tradeoff Score), defined as the harmonic mean of accuracy for answerable instances (acc_a) and unanswerable instances where the model must select the "unanswerable" option to deflect (acc_u):

$$\text{ADTScore} = \frac{2 \times acc_a \times acc_u}{acc_a + acc_u}$$

ADTScore is robust to class imbalance of *answerable* and *non-answerable* instances and achieves the maximum performance when model accuracy is balanced across both subsets.

Results Table 4 shows that while all LLMs handle multi-hop questions well, performance on time-

		Answerable		*Unanswerable*			
Model	**ADTScore**	**Multi H.**	**Time S.**	**Multi H.**	**Time S.**	**False P.**	**Uncertain S.**
Random	14.3	14.3	14.3	14.3	14.3	14.3	14.3
Phi3 mini (3.8B)	12.9	79.8	21.0	3.1	27.4	0.6	1.2
Phi3 small (7B)	23.5	80.5	34.8	10.2	39.5	8.2	4.5
Phi3 medium (14B)	19.9	79.7	53.8	16.6	17.1	8.3	5.5
Phi3.5 MoE (16×3.8B)	32.4	**82.9**	44.7	16.5	54.3	13.4	6.7
Qwen2.5 (7B)	31.5	67.5	42.9	26.7	23.1	16.2	21.8
Qwen2.5 (14B)	51.6	76.3	66.0	**44.9**	41.0	**42.8**	**32.7**
Qwen2.5 (32B)	**53.2**	79.4	67.3	41.7	**62.1**	38.6	26.7
Phi3 mini (3.8B) + CoT	41.9	53.8	29.1	42.7	43.8	41.9	37.7
Phi3 small (7B) + CoT	26.6	81.6	40.8	11.4	47.3	8.4	5.0
Phi3 medium (14B) + CoT	24.6	72.9	56.0	18.7	15.9	14.1	12.2
Phi3.5 MoE (16×3.8B) + CoT	38.0	78.1	46.1	23.3	54.4	19.4	15.5
Qwen2.5 (7B) + CoT	31.3	63.4	38.9	24.9	24.3	21.8	18.7
Qwen2.5 (14B) + CoT	49.6	75.8	66.2	42.0	41.9	39.3	29.3
Qwen2.5 (32B) + CoT	48.2	82.7	**69.5**	36.8	55.7	31.8	19.4

Table 4: NeoQA evaluation results for multi-hop, time-span, false premise, and uncertain specificity questions with all evidence up to the question date with/without CoT prompts. Metrics: ADTScore and accuracy by question type.

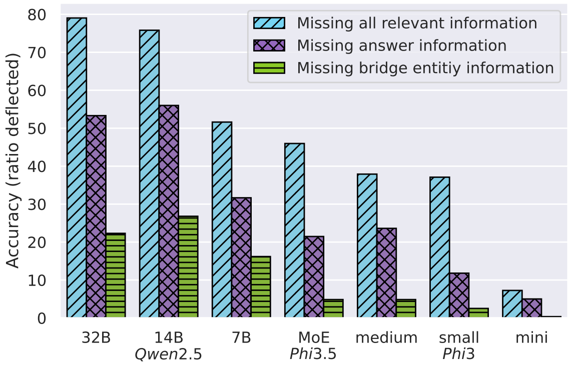


Figure 5: Model deflection ratio in multi-hop questions with varying evidence gaps.

span questions improves with model size. Unanswerable questions where the model must deflect are most challenging, especially for Phi3-based models. Qwen2.5 with 32B performs best overall with and ADTScore of 53.2 but still struggles to detect the subtle inconsistencies with false premise and uncertain specificity questions. Similar to prior work (Levy et al., 2024) we observe mixed effects on long contexts with CoT prompting, which mostly improves the performance of Phi3-based models by increasing their deflections, which boosts the overall score but harms multi-hop performance. The high answer parsing rate (97.9% with CoT, 99.2% without) suggests that mistakes arise from reasoning errors rather than poor instruction-following (see Appendix F.7).

Insufficient Evidence Multi-hop questions can lack sufficient evidence in three ways: (a) missing answer information (purple in Figure 2), (b) missing bridge entity information (green), or (c) missing both. The third case (c) occurred in 124 instances where both required information pieces were in the same article and removed together. When the answer itself was missing (cases a & c), the task resembled the IDK task by Vodrahalli et al. (2024). However, missing only the bridge entity (b) was the most challenging. In these cases, models often inferred the correct answer through shortcuts rather than recognizing the evidence as incomplete and deflecting appropriately. Figure 5 shows performance by evidence type (in all cases the distracting documents up to the question date remained). Models found it easiest to deflect when all relevant information was absent but struggled most when the answer was present while the bridge entity was missing. The errors followed a consistent pattern across models. When the answer itself was missing, they were more likely to select a misleading option (52.9%-77.9%). When the bridge entity was missing, models frequently answered as if nothing were missing (69.7%-90.7% of errors). In general, we observed that LLMs tend to overlook subtle differences between questions and evidence. Except for Phi3 (mini & small), which struggled with deflection, we found a significant negative association between accuracy on multi-hop questions with sufficient evidence and questions where bridge entity information was omitted or where questions were manipulated into false premise or uncertain specificity (phi coefficient: $\phi = -0.114$ to -0.374, $p < 0.001$). For details, see Appendix F.2 and F.3.

GPT-4 Turbo To compare with our RealTimeQA experiments in Section 2, we tasked GPT-4 Turbo with answering questions without evidence. We

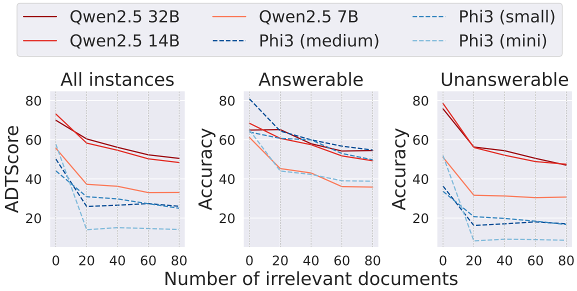


Figure 6: Performance over all instances (left), answerable (center), and unanswerable (right) instances with increasing number of irrelevant documents.

randomly sampled 250 multi-hop and time-span questions and converted them into four-way multiple choice format (excluding the deflection option). Accuracy was near random for time-span questions (24.4%) but higher for multi-hop questions (53.6%). Upon manual inspection of correctly predicted questions without evidence, we found no obvious give-away information in the question or answer options, and hypothesize that this is due to the synthetic data generation (see Appendix F.4 for examples). A model may benefit from learned token probabilities during inference because the dataset was sampled from the same token distribution. As discussed in Section 4.1, a clean separation between fictional and real-world knowledge is unrealistic. This highlights the importance of controlled experiments where reliance on parametric knowledge is penalized. We did not evaluate GPT-4 Turbo on the full dataset due to high computational costs and its large context requirements, especially since the data was generated by GPT-4o. For informativeness, we estimated its performance on 499 randomly sampled instances matching the question type distribution in Table 4. GPT-4 Turbo achieved an ADTScore of 42.4. It performed well on answerable questions (88% for multi-hop, 84% for time-span) but struggled with unanswerable ones—scoring just 25.3% on multi-hop with insufficient evidence, and 15% on both false premise and uncertain specificity questions. An exception was time-span questions with insufficient evidence, where it reached 58% accuracy.

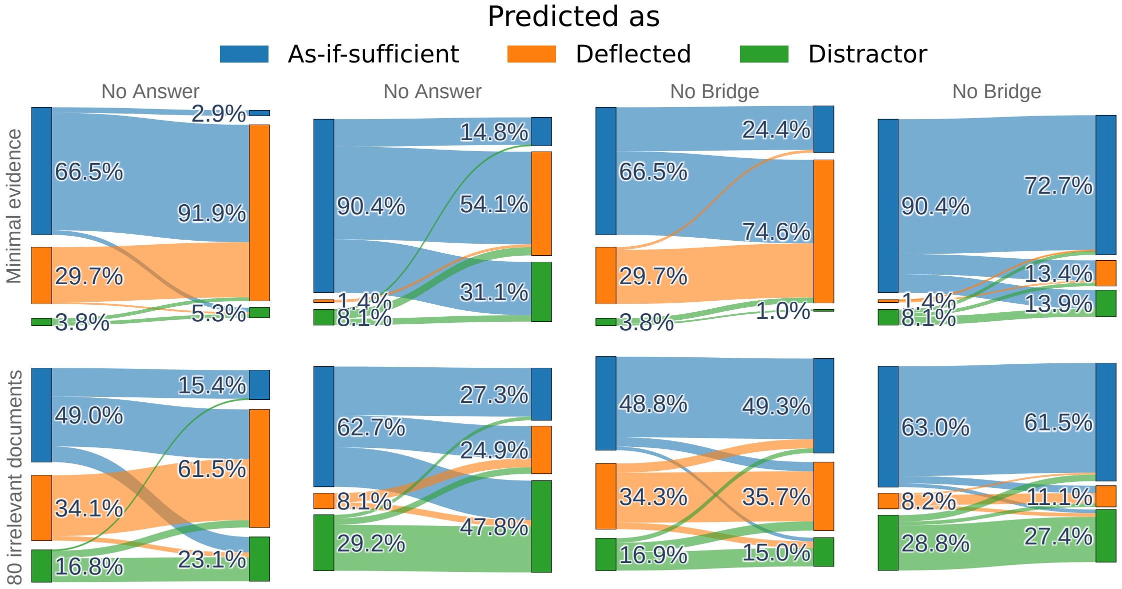
6 Impact of Irrelevant Documents

To evaluate the effect of irrelevant documents, we reuse the same questions—with sufficient and insufficient evidence, and unanswerable cases—and vary the number of irrelevant news articles from 0 to 80 in increments of 20. This results in 10,210 instances (Appendix E.3). Figure 6 shows the overall ADTScore and the aggregated accuracy for answerable and unanswerable instances. Across all models and configurations, performance is best when no irrelevant documents are present and declines as irrelevant documents are added. For answerable questions with only relevant evidence, all models perform in similar ranges. The smallest models in our experiments (Phi3 mini and Qwen2.5 7B) experience the sharpest decline, while the larger models are more robust. For each model, the major drop in performance occurs within the first 20 added irrelevant documents and then stabilizes, with the two larger Qwen2.5 models performing best. See Appendix F.5 for visualizations per model.

Prediction changes for insufficient evidence Figure 7 shows how multi-hop question predictions change when either answer or bridge entity information is omitted. The ideal model always predicts the correct answer (left , blue), when all evidence is available, and deflects (right, orange), when not. Without distracting news articles (top), Qwen2.5 14B mostly deflects with insufficient evidence, while Phi3 (medium) often selects distractors when the answer is missing or maintains its original prediction when bridge entity information is absent. Adding 80 irrelevant articles (bottom) decreases performance, but trends remain similar when answer information is omitted. When bridge entity information is missing, Qwen2.5 14B also predicts as though evidence were sufficient, highlighting the challenge of detecting insufficient evidence when distracting documents are present. Appendix F.6 shows the prediction changes for false premise and uncertain specificity questions.

7 Related Work

Several recent studies addressed parametric knowledge interference through time-sensitive questions (Vu et al., 2024; Yang et al., 2024b), dataset updates (Kasai et al., 2024), or automatic dataset generation from (recent) real-world data (Liska et al., 2022; Tang and Yang, 2024; Guinet et al., 2024). However, updated datasets introduce different instances, making direct performance comparisons over time unreliable, as fluctuations may result from dataset variations from different times (Luu et al., 2022). Other approaches focus on time-invariant questions (Wei et al., 2024) or conflicts between external and parametric knowledge (Longpre et al., 2021; Neeman et al., 2023; Tan et al.,



(a) **Qwen2.5 14B** (b) **Phi3 (med.)** (c) **Qwen2.5 14B** (d) **Phi3 (med.)**

Figure 7: Multi-hop question predictions change after removing key information containing the answer (*"No Answer"*) or the bridge entity (*"No Bridge"*). When evidence is sufficient (**left** in each diagram), the model must always predict the **correct answer**. When evidence is insufficient (**right** in each diagram), the model must **deflect**.

2024; Xu et al., 2024). In contrast, NeoQA generates a self-contained world, independent of real-world events and entities, to provide a stable testbed despite parametric knowledge updates. Our work extends needle-in-the-haystack tasks, where models must locate and reason over information in long text (Kamradt, 2023; Levy et al., 2024; Kuratov et al., 2024). These benchmarks are often based on real-world information or literature (Shaham et al., 2022, 2023; Bai et al., 2024; An et al., 2024; Liu et al., 2024; Wang et al., 2025; Hilgert et al., 2024), where parametric knowledge can interfere. Some mitigate this by constructing datasets from recent information (Li et al., 2024a; Karpinska et al., 2024). Apart from RGB (Chen et al., 2024), which heuristically determines relevant evidence, these datasets focus solely on answerable questions. Closest to our work is Michelangelo (Vodrahalli et al., 2024), which evaluates LLMs' long-context abilities using synthetic data outside their pretraining set and includes IDK questions where the answer is not in the text. NeoQA goes further by generating parallel worlds with recurring named entities and unanswerable questions with subtle mismatches with the grounding, rather than merely omitting explicit answers. Unanswerable questions have been studied in the context of adversarial manipulation (Rajpurkar et al., 2018; Sulem et al., 2021; Gautam et al., 2023), missing information in multi-hop reasoning (Trivedi et al., 2020, 2022; Atanasova et al., 2022), and false premises based on incorrect assumptions (Kim et al., 2021; Yu et al., 2023; Hu et al., 2023; Yang et al., 2024b). Similarly, we explore these challenges but focus on external evidence in a parallel world, where parametric knowledge can not detect flawed assumptions nor compensate for imperfect evidence.

8 Conclusion and Future Work

We introduce NeoQA, a novel dataset featuring out-of-training event timelines and question-answer pairs that are independent of real-world events. NeoQA serves as a robust platform for evaluating evidence-based question answering, as it requires models to answer questions exclusively from evidence and only when sufficient evidence is available. By automatically pairing questions and news articles, NeoQA simulates various retrieval conditions, ranging from scenarios with sufficient evidence to those with insufficient or irrelevant evidence. Our experiments across seven models reveal significant challenges in evidence-based reasoning: when key evidence required to answer a question is missing, models frequently resort to shortcut reasoning, a critical shortcoming for trustworthy applications. Future work may expand NeoQA with new questions and develop trustworthy models that reliably perform evidence-based reasoning.

Limitations

GPT-4-generated event timelines may reflect the LLM's social biases (Shin et al., 2024) and prompt-induced biases, making them unrepresentative of all real-world events. The challenges across question types depend on the LLM and prompts used. While suitable for zero-shot experiments, NeoQA is not appropriate for fine-tuning, as models could overfit to the generated question characteristics. LLMs may also introduce numerous, often intractable inconsistencies within timelines (Yang et al., 2022). Our experiments do not consider such possible inconsistencies. Our experiments are limited to the reported Phi3 and Qwen2.5 models due to licensing and legal restrictions, institutional policies, and the requirements of long-context LLMs. Due to these restrictions, we have not experimented with larger models and our findings are restricted to the evaluated models and sizes. Our work penalizes shortcut reasoning, which undermines model trustworthiness. However, some view it as beneficial for efficiency by skipping reasoning steps in CoT prompting (Ding et al., 2024). NeoQA is limited to English. Measuring the "naturalness" of generated text in a human study is challenging because the news articles produced by GPT-4o feature fictional entities, making it impossible to objectively compare them with real news. While we instructed GPT-4o to generate news articles in natural language and various styles, our manual evaluation focuses on the validity of the generated content rather than its naturalness.

Ethics Statement

All timelines and named entities are entirely fictional as approximated via Wikipedia, yet may include real-world entities if the LLM failed to detect named entities as such. The generated dataset may exhibit social biases, influenced by underlying social biases of LLMs. While our work is focused on creating fictional timelines, some events may unintentionally resemble real-world occurrences or entities. We emphasize that this data is fictional, and any similarities to real events or entities are purely coincidental and should be interpreted as such. Our paper passed an extensive multi-phase in-house review that took legal and ethical considerations into account. The human annotations in this paper are provided by qualified Mechanical Turk workers. We provided fair pay to our annotators. For the multi-hop answer evaluation, the workers take on average five minutes to complete one evaluation. We pay the workers \$0.35 per question and \$1.35 bonus, which leads to a pay of \$20.4 per hour.

Acknowledgments

The authors wish to thank Mengwen Liu for her insightful discussions and Andrei Coman, Adrià de Gispert, Weiwei Cheng and Saar Kuzi for their valuable feedback on an earlier version of this work. Max Glockner has been supported by the German Federal Ministry of Education and Research and the Hessian Ministry of Higher Education, Research, Science and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, and 1 others. 2024. Phi-4 Technical Report. *arXiv preprint arXiv:2412.08905*.

Chenxin An, Shansan Gong, Ming Zhong, Xingjian Zhao, Mukai Li, Jun Zhang, Lingpeng Kong, and Xipeng Qiu. 2024. L-Eval: Instituting Standardized Evaluation for Long Context Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14388–14411, Bangkok, Thailand. Association for Computational Linguistics.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2022. Fact Checking with Insufficient Evidence. *Transactions of the Association for Computational Linguistics*, 10:746–763.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2024. LongBench: A Bilingual, Multitask Benchmark for Long Context Understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3119–3137, Bangkok, Thailand. Association for Computational Linguistics.

Lang Cao. 2024. Learn to Refuse: Making Large Language Models More Controllable and Reliable through Knowledge Scope Limitation and Refusal Mechanism. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3628–3646, Miami, Florida, USA. Association for Computational Linguistics.

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.

Jifan Chen and Greg Durrett. 2019. Understanding Dataset Design Choices for Multi-hop Reasoning. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4026–4032, Minneapolis, Minnesota. Association for Computational Linguistics.

Jeffrey Cheng, Marc Marone, Orion Weller, Dawn Lawrie, Daniel Khashabi, and Benjamin Van Durme. 2024. Dated Data: Tracing Knowledge Cutoffs in Large Language Models. In *First Conference on Language Modeling*.

Yejin Choi. 2022. The Curious Case of Commonsense Intelligence. *Daedalus*, 151(2):139–155.

Elizabeth Clark, Yangfeng Ji, and Noah A. Smith. 2018. Neural Text Generation in Stories Using Entity Representations as Context. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2250–2260, New Orleans, Louisiana. Association for Computational Linguistics.

Tri Dao. 2024. FlashAttention-2: Faster Attention with Better Parallelism and Work Partitioning. In *The Twelfth International Conference on Learning Representations*.

Mengru Ding, Hanmeng Liu, Zhizhang Fu, Jian Song, Wenbo Xie, and Yue Zhang. 2024. Break the Chain: Large Language Models Can be Shortcut Reasoners. *arXiv preprint arXiv:2406.06580*.

Yanai Elazar, Akshita Bhagia, Ian Helgi Magnusson, Abhilasha Ravichander, Dustin Schwenk, Alane Suhr, Evan Pete Walsh, Dirk Groeneveld, Luca Soldaini, Sameer Singh, Hannaneh Hajishirzi, Noah A. Smith, and Jesse Dodge. 2024. What's In My Big Data? In *The Twelfth International Conference on Learning Representations*.

Lisa Fan, Marshall White, Eva Sharma, Ruisi Su, Prafulla Kumar Choubey, Ruihong Huang, and Lu Wang. 2019. In Plain Sight: Media Bias Through the Lens of Factual Reporting. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6343–6349, Hong Kong, China. Association for Computational Linguistics.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, and 7 others. 2020. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1307–1323, Online. Association for Computational Linguistics.

Vagrant Gautam, Miaoran Zhang, and Dietrich Klakow. 2023. A Lightweight Method to Generate Unanswerable Questions in English. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7349–7360, Singapore. Association for Computational Linguistics.

Max Glockner, Ieva Staliūnaitė, James Thorne, Gisela Vallejo, Andreas Vlachos, and Iryna Gurevych. 2024. AmbiFC: Fact-Checking Ambiguous Claims with Evidence. *Transactions of the Association for Computational Linguistics*, 12:1–18.

Zhibin Gou, Zhihong Shao, Yeyun Gong, yelong shen, Yujiu Yang, Nan Duan, and Weizhu Chen. 2024. CRITIC: Large Language Models Can Self-Correct with Tool-Interactive Critiquing. In *The Twelfth International Conference on Learning Representations*.

Gauthier Guinet, Behrooz Omidvar-Tehrani, Anoop Deoras, and Laurent Callot. 2024. Automated evaluation of retrieval-augmented language models with task-specific exam generation. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Lukas Hilgert, Danni Liu, and Jan Niehues. 2024. Evaluating and Training Long-Context Large Language Models for Question Answering on Scientific Papers. In *Proceedings of the 1st Workshop on Customizable NLP: Progress and Challenges in Customizing NLP for a Domain, Application, Group, or Individual (CustomNLP4U)*, pages 220–236, Miami, Florida, USA. Association for Computational Linguistics.

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. 2022. TRUE: Re-evaluating Factual Consistency Evaluation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3905–3920, Seattle, United States. Association for Computational Linguistics.

Shengding Hu, Yifan Luo, Huadong Wang, Xingyi Cheng, Zhiyuan Liu, and Maosong Sun. 2023. Won't Get Fooled Again: Answering Questions with False Premises. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5626–5643, Toronto, Canada. Association for Computational Linguistics.

Alon Jacovi, Avi Caciularu, Omer Goldman, and Yoav Goldberg. 2023. Stop Uploading Test Data in Plain Text: Practical Strategies for Mitigating Data Contamination by Evaluation Benchmarks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5075–5084, Singapore. Association for Computational Linguistics.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, 55(12):1–38.

Yichen Jiang and Mohit Bansal. 2019. Avoiding Reasoning Shortcuts: Adversarial Evaluation, Training, and Model Development for Multi-Hop QA. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2726–2736, Florence, Italy. Association for Computational Linguistics.

G. Kamradt. 2023. LLMTest: Needle In A Haystack. `https://github.com/gkamradt/LLMTest_NeedleInAHaystack`.

Marzena Karpinska, Katherine Thai, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2024. One Thousand and One Pairs: A "novel" challenge for long-context language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17048–17085, Miami, Florida, USA. Association for Computational Linguistics.

Jungo Kasai, Keisuke Sakaguchi, Ronan Le Bras, Akari Asai, Xinyan Yu, Dragomir Radev, Noah A Smith, Yejin Choi, Kentaro Inui, and 1 others. 2024. REALTIME QA: what's the answer right now? *Advances in Neural Information Processing Systems*, 36.

Norambuena Keith, Tanushree Mitra, and Chris North. 2023. A Survey on Event-Based News Narrative Extraction. *ACM Computing Surveys*, 55(14s):1–39.

Najoung Kim, Ellie Pavlick, Burcu Karagol Ayan, and Deepak Ramachandran. 2021. Which Linguist Invented the Lightbulb? Presupposition Verification for Question-Answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3932–3945, Online. Association for Computational Linguistics.

Yuri Kuratov, Aydar Bulatov, Petr Anokhin, Ivan Rodkin, Dmitry Igorevich Sorokin, Artyom Sorokin, and Mikhail Burtsev. 2024. BABILong: Testing the limits of LLMs with long context reasoning-in-a-haystack. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Yukyung Lee, Soonwon Ka, Bokyung Son, Pilsung Kang, and Jaewook Kang. 2025. Navigating the Path of Writing: Outline-guided Text Generation with Large Language Models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track)*, pages 233–250, Albuquerque, New Mexico. Association for Computational Linguistics.

Mosh Levy, Alon Jacoby, and Yoav Goldberg. 2024. Same Task, More Tokens: the Impact of Input Length on the Reasoning Performance of Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15339–15353, Bangkok, Thailand. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Jiaqi Li, Mengmeng Wang, Zilong Zheng, and Muhan Zhang. 2024a. LooGLE: Can Long-Context Language Models Understand Long Contexts? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long*

Papers), pages 16304–16333, Bangkok, Thailand. Association for Computational Linguistics.

Yifei Li, Xiang Yue, Zeyi Liao, and Huan Sun. 2024b. AttributionBench: How Hard is Automatic Attribution Evaluation? In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14919–14935, Bangkok, Thailand. Association for Computational Linguistics.

Xiao Ling and Daniel Weld. 2021. Fine-Grained Entity Recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 26(1):94–100.

Adam Liska, Tomas Kocisky, Elena Gribovskaya, Tayfun Terzi, Eren Sezener, Devang Agrawal, D'Autume Cyprien De Masson, Tim Scholtes, Manzil Zaheer, Susannah Young, and 1 others. 2022. StreamingQA: A Benchmark for Adaptation to New Knowledge over Time in Question Answering Models. In *International Conference on Machine Learning*, pages 13604–13622. PMLR.

Bang Liu, Di Niu, Kunfeng Lai, Linglong Kong, and Yu Xu. 2017. Growing story forest online from massive breaking news. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 777–785.

Nelson Liu, Tianyi Zhang, and Percy Liang. 2023. Evaluating Verifiability in Generative Search Engines. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7001–7025, Singapore. Association for Computational Linguistics.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. Entity-Based Knowledge Conflicts in Question Answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7052–7063, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kelvin Luu, Daniel Khashabi, Suchin Gururangan, Karishma Mandyam, and Noah A. Smith. 2022. Time Waits for No One! Analysis and Challenges of Temporal Misalignment. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5944–5958, Seattle, United States. Association for Computational Linguistics.

Yan Ma, Yu Qiao, and Pengfei Liu. 2024. MoPS: Modular Story Premise Synthesis for Open-Ended Automatic Story Generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2135–2169, Bangkok, Thailand. Association for Computational Linguistics.

Inbal Magar and Roy Schwartz. 2022. Data Contamination: From Memorization to Exploitation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 157–165, Dublin, Ireland. Association for Computational Linguistics.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. AmbigQA: Answering Ambiguous Open-domain Questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5783–5797, Online. Association for Computational Linguistics.

Ella Neeman, Roee Aharoni, Or Honovich, Leshem Choshen, Idan Szpektor, and Omri Abend. 2023. DisentQA: Disentangling Parametric and Contextual Knowledge with Counterfactual Question Answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10056–10070, Toronto, Canada. Association for Computational Linguistics.

Adithya Pratapa, Kevin Small, and Markus Dreyer. 2023. Background Summarization of Event Timelines. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8111–8136, Singapore. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Oscar Sainz, Iker García-Ferrero, Alon Jacovi, Jon Ander Campos, Yanai Elazar, Eneko Agirre, Yoav Goldberg, Wei-Lin Chen, Jenny Chim, Leshem Choshen, Luca D'Amico-Wong, Melissa Dell, Run-Ze Fan, Shahriar Golchin, Yucheng Li, Pengfei Liu, Bhavish Pahwa, Ameya Prabhu, Suryansh Sharma, and 9 others. 2024. Data Contamination Report from the 2024 CONDA Shared Task. In *Proceedings of the 1st Workshop on Data Contamination (CONDA)*, pages 41–56, Bangkok, Thailand. Association for Computational Linguistics.

Tal Schuster, Adam Fisch, and Regina Barzilay. 2021. Get your vitamin C! robust fact verification with contrastive evidence. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 624–643, Online. Association for Computational Linguistics.

Tal Schuster, Darsh Shah, Yun Jie Serene Yeo, Daniel Roberto Filizzola Ortiz, Enrico Santus, and Regina Barzilay. 2019. Towards Debiasing Fact Verification Models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3419–3425, Hong Kong, China. Association for Computational Linguistics.

Dafna Shahaf and Carlos Guestrin. 2010. Connecting the dots between news articles. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 623–632.

Uri Shaham, Maor Ivgi, Avia Efrat, Jonathan Berant, and Omer Levy. 2023. ZeroSCROLLS: A Zero-Shot Benchmark for Long Text Understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7977–7989, Singapore. Association for Computational Linguistics.

Uri Shaham, Elad Segal, Maor Ivgi, Avia Efrat, Ori Yoran, Adi Haviv, Ankit Gupta, Wenhan Xiong, Mor Geva, Jonathan Berant, and Omer Levy. 2022. SCROLLS: Standardized CompaRison Over Long Language Sequences. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 12007–12021, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jisu Shin, Hoyun Song, Huije Lee, Soyeong Jeong, and Jong Park. 2024. Ask LLMs Directly, "What shapes your bias?": Measuring Social Bias in Large Language Models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 16122–16143, Bangkok, Thailand. Association for Computational Linguistics.

Aviv Slobodkin, Omer Goldman, Avi Caciularu, Ido Dagan, and Shauli Ravfogel. 2023. The Curious Case of Hallucinatory (Un)answerability: Finding Truths in the Hidden States of Over-Confident Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3607–3625, Singapore. Association for Computational Linguistics.

Elior Sulem, Jamaal Hay, and Dan Roth. 2021. Do We Know What We Don't Know? Studying Unanswerable Questions beyond SQuAD 2.0. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4543–4548, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hexiang Tan, Fei Sun, Wanli Yang, Yuanzhuo Wang, Qi Cao, and Xueqi Cheng. 2024. Blinded by Generated Contexts: How Language Models Merge Generated and Retrieved Contexts When Knowledge Conflicts? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6207–6227, Bangkok, Thailand. Association for Computational Linguistics.

Yixuan Tang and Yi Yang. 2024. MultiHop-RAG: Benchmarking Retrieval-Augmented Generation for Multi-Hop Queries. In *First Conference on Language Modeling*.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2020. Is Multihop QA in DiRe Condition? Measuring and Reducing Disconnected Reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8846–8863, Online. Association for Computational Linguistics.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. MuSiQue: Multihop Questions via Single-hop Question Composition. *Transactions of the Association for Computational Linguistics*, 10:539–554.

Kiran Vodrahalli, Santiago Ontanon, Nilesh Tripuraneni, Kelvin Xu, Sanil Jain, Rakesh Shivanna, Jeffrey Hui, Nishanth Dikkala, Mehran Kazemi, Bahare Fatemi, and 1 others. 2024. Michelangelo: Long context evaluations beyond haystacks via latent structure queries. *arXiv preprint arXiv:2409.12640*.

Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, and Thang Luong. 2024. FreshLLMs: Refreshing Large Language Models with Search Engine Augmentation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13697–13720, Bangkok, Thailand. Association for Computational Linguistics.

Cunxiang Wang, Ruoxi Ning, Boqi Pan, Tonghui Wu, Qipeng Guo, Cheng Deng, Guangsheng Bao, Xiangkun Hu, Zheng Zhang, Qian Wang, and Yue Zhang. 2025. NovelQA: Benchmarking Question Answering on Documents Exceeding 200K Tokens. In *The Thirteenth International Conference on Learning Representations*.

Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. 2024. Measuring short-form factuality in large language models. *arXiv preprint arXiv:2411.04368*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *Advances in neural information processing systems*, 35:24824–24837.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. 2024. Knowledge conflicts for LLMs: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8541–8565, Miami, Florida, USA. Association for Computational Linguistics.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024a. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Kevin Yang, Dan Klein, Nanyun Peng, and Yuandong Tian. 2023. DOC: Improving Long Story Coherence With Detailed Outline Control. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3378–3465, Toronto, Canada. Association for Computational Linguistics.

Kevin Yang, Yuandong Tian, Nanyun Peng, and Dan Klein. 2022. Re3: Generating longer stories with recursive reprompting and revision. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4393–4479, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Xiao Yang, Kai Sun, Hao Xin, Yushi Sun, Nikita Bhalla, Xiangsen Chen, Sajal Choudhary, Rongze Daniel Gui, Ziran Will Jiang, Ziyu Jiang, Lingkun Kong, Brian Moran, Jiaqi Wang, Yifan Ethan Xu, An Yan, Chenyu Yang, Eting Yuan, Hanwen Zha, Nan Tang, and 8 others. 2024b. CRAG - Comprehensive RAG Benchmark. In *Advances in Neural Information Processing Systems*, volume 37, pages 10470–10490. Curran Associates, Inc.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Hao Yu, Aoran Gan, Kai Zhang, Shiwei Tong, Qi Liu, and Zhaofeng Liu. 2025. Evaluation of Retrieval-Augmented Generation: A Survey. In *Big Data*, pages 102–120, Singapore. Springer Nature Singapore.

Xinyan Yu, Sewon Min, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. CREPE: Open-domain question answering with false presuppositions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10457–10480, Toronto, Canada. Association for Computational Linguistics.

Xiang Yue, Boshi Wang, Ziru Chen, Kai Zhang, Yu Su, and Huan Sun. 2023. Automatic evaluation of attribution by large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4615–4635, Singapore. Association for Computational Linguistics.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2023. Large language models are human-level prompt engineers. In *The Eleventh International Conference on Learning Representations*.

Hanlin Zhu, Andrew Cohen, Danqing Wang, Kevin Yang, Xiaomeng Yang, Jiantao Jiao, and Yuandong Tian. 2023. End-to-end Story Plot Generator. *arXiv preprint arXiv:2310.08796*.

A RealtimeQA Experiments

The RealtimeQA dataset spans weekly news quizzes from June 16, 2022, to January 12, 2024. We select 1,548 questions with four answer options where gold evidence is provided, enabling direct comparison of model performance using evidence versus parametric knowledge. Our experiments use GPT-4 Turbo ("gpt-4-turbo-2024-04-09") with the reported knowledge cutoff in December, 2023. To include sufficient questions beyond the reported cutoffs, we collected 660 additional instances from January 18 to September 13, 2024, via the Wayback Machine[3] using the same sources as RealtimeQA, totaling 2,208 questions. It is important to emphasize that this experiment cannot definitively attribute the observed decrease in performance to reduced parametric knowledge alone, as it may also result from variations in the instances themselves. Nonetheless, the consistent downward trend across all models over time strongly suggests that the decline is primarily due to less useful parametric knowledge.

B Timeline Example

The complete outline of the first event from Figure 2 is shown in Figure 8. Each outline item has a unique ID within the timeline and conveys distinct, specific information. Table 5 shows the updated KB entry after the final event for the fictional person Selvia Renek. Figure 9 shows a news article with resolved named entities. Figure 10 shows an complete instance with a question, answer options and news articles as evidence.

C Dataset Generation

C.1 Timeline

C.1.1 Initial Summary Generation

Automatic story premise generation follows narrative templates, like protagonist and antagonist (Ma et al., 2024), which differs from real-world event progression. Therefore, we base timeline generation on *event summaries*, generated automatically. Specifically, we generate diverse initial summaries for the timelines in three steps. First, using GPT-4, we generated 20 news genres:

- Art
- Business
- Celebrities
- Crimes
- Economics
- Education
- Environment
- Epidemics
- Food
- Health
- International Affairs
- Legal
- Lifestyle
- Local News
- Politics
- Science
- Social Issues
- Sports
- Technology
- Travel

While these genres may overlap, they create diversity in the direction of initial event summaries. Second, for each genre, the LLM generates 20 different generic event types. Examples include:

- Retirement Living and Senior Lifestyle Changes (genre: **Lifestyle**)
- Gourmet Food and Culinary Experiences (genre: **Lifestyle**)
- Scandals and Controversies (genre: **Celebrities**)
- Celebrity Weddings (genre: **Celebrities**)
- Major Tournaments Outcomes (genre: **Sport**)
- Player Transfer and Trades (genre: **Sport**)

Third, for each generic event type and genre, the LLM generates ten different event summaries without using named entities. For example, in the "Gourmet Food and Culinary Experiences" genre "Lifestyle", the summaries include:

- A coastal city is set to host its first-ever seafood festival, featuring sustainable fishing practices and cooking demonstrations by renowned chefs.
- A culinary school has announced a new program focusing on the art of fermentation, aiming to teach techniques from around the world.
- Experts in plant-based cuisine have gathered for a conference to explore the future of vegan gourmet food, sharing innovations in texture and flavor.
- A pop-up restaurant specializing in fusion desserts has opened for a limited time, offering a blend of traditional and modern sweets from various cultures.

[3] https://web.archive.org/

ID	PERSON-3
Name	Selvia Renek
Entity Class	Person
Description	Selvia Renek is an enthusiastic and detail-oriented food critic who focuses on celebrating local culinary creativity.
Date of Birth	1992-03-09
Gender	Female
Profession	Food Blogger
Nationality	Varentian
Education	Bachelors in Journalism, University of Alveris
Height	160 cm
Hair Color	Auburn
Eye Color	Blue
Affiliation	Progressive
Marital Status	Single
	History
2024-03-15	Praised Amber Silk as "the most delicate balance of flavors" and contributed to the event's buzz on social media.
2024-04-10	Published an investigative blog post highlighting unethical practices in Amber Glaze Delights' supply chain.
2024-07-03	Published an op-ed discussing implications of Amber Glaze Delights' audit findings and emphasized systemic change in sourcing oversight.
2024-07-08	Published an article questioning the robustness of ethical certification programs and called for increased community participation in refining criteria.
2024-07-09	Published a blog spotlighting Amara Hearth Café's plans to adopt ethical sourcing practices, including an interview with Erena Treflin.
2024-07-10	Published a SnapGram post summarizing the forum, highlighting tensions between ethical practices and accessibility for smaller businesses.
2024-07-21	Hosted a live SnapGram Q&A session to address community concerns about the Calder Square Cultural Committee's ethical certification program.
2024-08-01	Published a blog detailing the rollout of the self-certification program and included interviews with representatives and vendors.

Table 5: The updated KB entry for the fictional person Selvia Renek after the last event of the timeline. For each event in which Selvia Renek participated, the history summarizes her role.

From these summaries, we randomly sample 15 to create NeoQA, while the remaining summaries are published for future work. The last event summary resulted in the timeline shown in Figure 2.

C.1.2 Timeline Generation

The LLM uses 12 prompts to generate the timeline. Each prompt incorporates critiques to detect and correct errors in real-time. The process involves: *i)* generating the event outline, *ii)* creating and updating fictional named entities, and *iii)* generating summaries as seed for the next event. Each outline includes at least 20 detailed, date-specific sentences as outline items.

Step 1: Event Outline Generation Given a seed summary, all previously generated outlines, and named entity KB entries from the same timeline, the LLM generates a new 10-sentence outline, with each sentence (outline item) capturing one aspect of the fictional event. A temperature of $t = 1.0$ is used to promote creative story progression (Prompt H.2).

Step 2: Event Outline Refinement Given the generated outline, along with all previously generated event outlines and named entity KB entries, the LLM refines each outline item in the outline by adding up to two additional outline items that provide specific details about the original outline item. This ensures a highly detailed event outline while keeping its scope constrained by the outline item count defined in the previous step. A temperature of $t = 1.0$ is used to enhance creativity (Prompt H.3).

Step 3: Outline Consistency To address inconsistencies from the high-temperature generation of the initial outline, the LLM checks and corrects the consistency of the outline with previous event outlines and named entity KB entries. This step uses a temperature of $t = 0.0$ for deterministic output, with critiques ensuring the outline item count remains unchanged (Prompt H.4).

Step 4: Named Entity Recognition (novel named entities) Given the outline and all previously generated named entity KB entries, the LLM detects each novel named entity in the outline. We heuristically verify that the identified entities do not overlap with existing entries from previous events and restrict each entity to a maximum of five words. The temperature is set to $t = 0.0$. Possible named entity types include Location, Person, Organization, Product, Art, Event, Building, and Miscellaneous (Prompt H.5).

Step 5: Fictional Named Entities Generation We compare each newly detected named entity with Wikipedia[4]. The API accounts for variations in names, such as "Obama" and "Barack Obama." If a match is found, we query the LLM to generate different fictional names that fit the outline's context, continuing until no search results for the named entity exist. We heuristically ensure that the LLM does not increase the entity's length beyond five words or use brackets, to avoid generating names that overlap significantly with existing entities. The temperature is set to $t = 1.0$ (Prompt H.6).

Step 6: Adjust the Outline The LLM refines the outline using the newly generated names for the named entities. We ensure that the previous names no longer appear in the outline. The temperature is set to $t = 0.0$ (Prompt H.7).

Step 7: Named Entity Recognition (all named entities) The LLM identifies all named entities in the outline, considering both new and existing entities. Although this step may overlap with previous ones, we have found that isolating this step helps minimize errors in named entity detection, which is crucial for event generation and consistency. We ensure that all novel named entities are included and that the LLM does not output unknown entities. The temperature is set to $t = 0.0$ (Prompt H.8).

Step 8: Named Entity Resolution in the Outline The LLM marks all named entities in the outline using the format `[phrase]|[entity-id]`, where `phrase` is the name as referenced in the outline. We heuristically verify that the outline does not contain unresolved named entities and that all detected named entities are properly resolved within the outline. The temperature is set to $t = 0.0$ (Prompt H.9).

Step 9: Populate New Named Entity KB Entries The LLM generates new KB entries for each new named entity, with each entry including a name and description. Different types of named entities have different additional fields. We cross-check the populated KB entry fields with Wikipedia and prompt the model to correct the entries until none of the properties reference known named entities according to Wikipedia. The temperature is set to $t = 1.0$ (Prompt H.10).

Step 10: Update Named Entity KB Entries Based on the outline, the LLM generates a single sentence for each named entity involved in the current event, describing the entity's role in the event and/or how the event affected it. The LLM may also update properties of the named entities, such as the budget of an "event" or the number of employees of an "organization". We cross-check the updated properties to ensure they are distinct from those in Wikipedia. The temperature is set to $t = 0.0$ (Prompt H.11).

Step 11: Generate Diverse Next Summaries The LLM generates a set of diverse summaries for the subsequent event. We prompt the LLM to create summaries with different story directions, varying impacts, and both positive and negative developments. One of the generated summaries is then randomly selected. The temperature is set to $t = 1.0$ (Prompt H.12).

Step 12: Generate Mutually Exclusive Summaries Given the selected summary for the next event, the LLM generates three mutually exclusive summaries, where only one can occur. This introduces irreversible story continuations that diverge from the most likely continuation based on the model's parametric knowledge (and thus aligned with past real-world events). The temperature is set to $t = 1.0$ (Prompt H.13).

Step 13: Next Event Generation Continue with **Step 1** using one randomly selected summary from the mutually exclusive summaries.

C.2 Questions

The question generation process consists of three phases:

1. **Outline Item Selection:** Given one or two events, the LLM selects a subset of two outline items as the basis for the question.

2. **Question Writing:** Using the selected outline items, the full event history, and named entities, the LLM drafts a question and answer pair that (a) can be answered using the selected outline items, (b) requires both outline items for a complete answer, and (c) has a

[4] `https://en.wikipedia.org/w/api.php`

unique, unambiguous answer based on all past context. If necessary, additional outline items from the selected events may be included.

3. **Distractor Generation:** The LLM generates plausible but incorrect distractor options for the question, along with justifications explaining their plausibility and incorrectness.

Each question requires two distinct pieces of information (or more, if additional outline items are added) for sufficient grounding. Multiple questions are generated for all $\binom{n+1}{2} = 55$ combinations of two events (including each event individually) in each timeline with $n = 10$ events. Across all steps, the temperature is set to $t = 0.0$.

C.2.1 Time-span Questions

Step 1: Select outline items The model selects pairs of outline items from two provided events outlines to identify time points for calculating a meaningful duration. The latter event defines the question's date. Critiques ensure the chosen outline items come from different events when multiple events are provided (Prompt H.14).

Step 2: Write question Using all event outlines up to the question date and the selected outline items, the model generates a question about the durations between two points in time, defined by the selected outline items. The LLM must ensure the selected outline items provide sufficient evidence to answer the question while requiring both. If additional context is needed, the LLM may add an outline item but must justify its inclusion (Prompt H.15).

Step 3: Refine answerability In initial iterations we observed that generated questions often relied on unnecessary or unstated assumptions. To address this, we instruct the LLM to evaluate each assumption for necessity, to ensure it does not contain relevant information from the selected outline items, and to add any missing assumptions if needed (Prompt H.16).

Step 3: Create distractors Given all previous event outlines, the generated question with the correct answer, and the selected outline items, the LLM generates 5 distractor answers. We instruct the LLM to make use of the content of the event outlines to craft challenging distractors and provide a rationale for each, explaining why it is misleading yet incorrect (Prompt H.17).

C.2.2 Multi-hop Questions

Step 1: Select outline items For each event combination, we identify named entities common to both events and randomly choose two. For each chosen entity, we prompt the LLM to select two outline items from the event outlines that can form a multiple-hop question with a bridge entity. Critiques are used to ensure the outline items discuss the selected entity (Prompt H.18).

Step 2: Write question Given the past event outlines and the two selected outline items discussing the same named entity, the LLM creates a multi-hop question with the correct answer. The question should ask about the bridge entity's information from one outline item while paraphrasing it using information from the other, as described in Yang et al. (2018). The question must have a unique and unambiguous answer based on the past event outlines (Prompt H.19).

Step 3: Create distractors Using the previous event outlines, generated question and answer, and selected outline items, the LLM creates plausible yet incorrect distractor answers. For each distractor, the LLM provides a justification explaining why it is incorrect but still plausible (Prompt H.20).

C.2.3 Unanswerable Questions

We create unanswerable questions by modifying the generated multi-hop questions, reusing the selected outline items and distractor options. Given the generated multi-hop question and the past event outlines, the LLM to makes subtle adjustments to the question to introduce contradictions for false-premise questions (Prompt H.21) or add additional constraints that cannot be confirmed or denied by the event outlines for uncertain specificity questions (Prompt H.22).

C.3 News Articles

C.4 Generation

We use GPT-4 to generate four distinct news profiles, each defining unique values, reporting style, perspective on common issues, preferred topics, and likes and dislikes.

1. SensationalNews (System Prompt H.23)
2. ObjectiveNews (System Prompt H.24)
3. ProgressiveNews (System Prompt H.25)
4. ConservativeNews (System Prompt H.26)

Generating a news article follows these steps:

Step 1: Select outline items Given the outlines of all past events, and the current event outline, the LLM selects four subsets of outline item ids from the current event outline, which will be used to generate four different news articles. We use a temperature of $t = 0.0$ and the news profile as system prompt (Prompt H.27).

Step 2: Write the news article Using basic information about the named entities (excluding event update histories) and the selected outline items, the LLM generates a news article with a headline that aligns with the selected outline items. Including basic information about the named entities helps contextualize their relationships (e.g., a person being the head of a company) and provides their full names. To ensure diversity, we set the temperature to $t = 1.0$ and use the news profile as the system prompt (Prompt H.28).

Step 3: Remove hallucinations Generating news articles with high temperature increases diversity but risks hallucinations (Ji et al., 2023) that diverge from the selected outline items. To mitigate this, the LLM removes unverifiable information while retaining the article's style. Unfaithful content is permitted only if clearly hedged as hypothetical. The LLM is prompted without a newspaper profile as system prompt and with a temperature of $t = 0.0$ (Prompt H.29).

Step 4: Add missing content To ensure all required information is conveyed, the LLM compares the selected outline items with the generated news article and ensures all details are included. As in step 3, the LLM is instructed to maintain the article's original style and is prompted without a newspaper profile as system prompt, using a temperature of $t = 0.0$ (Prompt H.30).

Step 5: Named entity resolution The LLM identifies and marks all named entities in the news article using the named entity KB entries. We use a temperature of $t = 0.0$ (Prompt H.31).

C.5 News article statistics

We generated a total of 1,800 news articles. The token count per article, measured using the tiktoken tokenizer (version 0.8.0) for the "GPT-4" model, ranges from 222 to 603 tokens, with an average length of 356.2 tokens (±47.7).[5]. Figure 11 shows the overall token distribution.

D Quality Measures

We evaluate NEOQA on three dimensions: *i*) whether the questions are answerable using the isolated outline items used to generate them, *ii*) whether the news articles convey the expected content, and *iii*) whether the question-evidence pairs are valid. The first two assess key assumptions for assembling questions with evidence documents based on the selected outline items, while the third evaluates answerability from the combined question and news article, independent of these assumptions.

D.1 Question Filtering

We filter questions in two steps: First, remove answerable (multi-hop and time-span) questions that cannot be answered using the selected outline items as evidence, along with unanswerable questions derived from them. Second, remove time-span questions that can be answered with fewer outline items than selected, as their temporal assumptions are explicitly stated in the question. Results are shown in Table 6. This step removed 41.8% of answerable questions (30.1% of multi-hop and 51.5% of time-span questions) along with unanswerable ones derived from them (30.0% of false premise and 30.5% of uncertain specificity questions).

Step	Multi H.	Time S.	False P.	Uncertain S.
(1)	1,201	1,438	4,114	4,245
(2)	−362	−564	−1,235	−1,293
(3)	±0	−196	±0	±0
Final	**839**	**678**	**2,879**	**2,952**

Table 6: Number of question types after initial generation **(1)**, filtering for answerability **(2)**, and for leaked assumptions **(3)**.

Answerability filtering During question generation, the LLM has access to the full outlines of prior events, not just the selected evidence outline items. This allows the LLM to prevent issues like ambiguous or time-sensitive answers that are not unique given the past evidence, but may introduce dependencies beyond the selected evidence, violating the assumption that these outline items alone are sufficient. Additionally, LLMs may make errors during question-answer generation. To improve NEOQA quality, we remove questions the LLM cannot correctly answer using only the selected evidence outline items (i.e., using perfect evidence; Prompts H.32 & H.33). If the LLM fails to answer

[5] https://github.com/openai/tiktoken

its own question with perfect evidence, we discard the question. This conservatively excludes questions where the LLM cannot reverse its reasoning to answer correctly under optimal conditions. Out of 1,201 multi-hop and 1,438 time-span questions, we discarded 362 (30.1%) and 564 (39.2%), respectively. Most questions (91.0%) were discarded because the LLM deemed the selected outline items insufficient. In only the remaining 9.0% of discarded questions, the LLM predicted a distractor instead of the assumed correct answer. This also led to the removal of 1,235 false-premise questions (out of 4,114) and 1,293 uncertain-specificity questions (out of 4,245) derived from the discarded multi-hop questions.

Leaked assumption filtering Outlines sometimes include vague temporal terms like "early June" making it difficult to generate duration-based questions that meet strict grounding requirements in our task definition. To address this, we instructed the LLM to define clear assumptions (e.g., "assume early June refers to June 1st") during question generation. However, manual review revealed questions where these assumptions replaced some or all of the evidence, with required details from the evidence restated in the question (see Table 7). To remove such questions, we conducted a second LLM-based filtering step. The LLM was tasked with answering using insufficient evidence (Prompt H.34). If it could derive the correct answer from any subset of insufficient evidence, we discarded the question. This step eliminated 22.4% (196) of the remaining time-span questions. Specifically, we removed 169 questions where the LLM could omit one required evidence outline item and 27 where it required no evidence at all.

D.2 NLI Verification

We use a pretrained LLM to test our assumption that each news article fully conveys the information in its selected outline items. Specifically, we employ the T5-XXL (Raffel et al., 2020), provided by Honovich et al. (2022), trained as a binary classifier on six NLI and fact-checking datasets. The model determines whether a premise text entails a hypothesis (output: "1") or not (output: "0"). The generated news article serves as the premise text, and each outline item from the outline acts as a hypothesis text. We expect the model to predict entailment ("1") for selected outline items included in the article and no-entailment ("0") for other event outline items. If the model outputs a label outside the expected ones, we assume no clear entailment and mark the prediction as incorrect by default.

Results We computed NLI predictions for every outline item of each event against every generated news article, resulting in 380,880 outline item-article pairs (Table 8). According to the NLI model the expected content is contained in the news articles in 98.1%. For outline items expected to be excluded, the model agreed in 92.2%. Notably, only for 0.5% of the outline items it directly disagreed our expected label and predicted "entailed" rather than "not-entailed". In most cases the outline items where predicted as "unknown" rather than "entailed". This observation was consistent across outline items from the same event and other events. We did not compute numbers for the "entailed" outline items separately, as this label applies only when the outline items and news article are from the same event.

D.3 Human Annotation

We use the Amazon Mechanical Turk (AMT) platform[6] to collect human annotations for evaluating NeoQA on a subset of 350 instances. Each question type is annotated separately. For answerable questions (multi-hop and time-span), annotators select the correct answer based on two sufficient news articles, choosing from one correct option, one distractor, and one option for unanswerable questions (see Figure 12 for an example). Preliminary annotations by the authors and by crowd workers revealed that unanswerable questions (false premise, uncertain specificity) are particularly challenging, as they require identifying relevant evidence and nuanced mismatches in detailed news articles. To simplify this annotation, we provide only the relevant outline items (two sentences) instead of complete news articles. Annotators then determine whether the question can be answered based on these sentences, selecting either the original multi-hop answer or the unanswerable option (see Figure 13). To help annotators focus on key information, we include LLM-generated justifications for each answer option, guiding their attention to the relevant information. Each justification begins with "This answer is correct because [...]." This approach mitigates cognitive load, reducing annotator fatigue and improving annotation quality, as was observed in preliminary annotations by the authors.

[6] https://www.mturk.com/

Validity	Question	Explanation
valid	Assuming that the six-month monitoring period for the updated implementation of the injury risk categorization tool begins on the date of its announcement, what is the time span between the conclusion of Aleena Karentov's motivational talk at the end of her session and the end of this monitoring period?	*The assumption reduces uncertainty from the evidence outline items but is meaningless if the relevant evidence is missing.*
partially leaked	What is the duration between the announcement of the compromise plan concerning the curriculum at the Murvenstad Gymnastics Alliance and the earliest possible start date of the follow-up workshops, assuming they begin on the earliest possible date in July 2026?	*The assumption defines the start date of the compromise plan, one of the two required points in time.*
fully leaked	Assuming the six-month performance metric collection period for the pilot community centers begins on 2027-03-01 and the second round of field tests by Stranlen Transport Solutions is planned to start on 2027-10-01, what is the time span between the end of the performance metric collection period and the start of the second round of field tests?	*The assumptions identify all required points in time, making the evidence document optional.*

Table 7: Time-span questions with assumptions that are required (*valid*) or those that leak critical information from the evidence (*partially leaked / leaked*).

			Predicted as		
Instances	**Expected Label**	**Count**	**Entailed**	**Not Entailed**	**Unknown**
All	entailed	10,675	**98.1%**	1.8%	0.1%
All	not entailed	370,205	0.5%	**92.2%**	7.3%
Same Event	not-entailed	27,413	0.9%	**95.0%**	4.1%
Different Event	not-entailed	342,792	0.5%	**92.0%**	7.5%

Table 8: NLI predictions between event outline items and news articles.

Annotation Each task posted on Mechanical Turk clearly described the nature of the work, and the compensation offered. The annotator must voluntarily accept the task to start working on each HIT. We protect the privacy and confidentiality of our annotators. We do not collect personal information from the AMT workers; each worker is identified by a unique ID. We followed Mechanical Turk's terms of use and guidelines, ensuring that our research did not violate any platform-specific rules. We restrict participation to a pre-selected pool of annotators with proven English proficiency and a history of high-quality annotations. Pay is set at $0.35 per question, plus a $1.35 bonus, resulting in an hourly rate of $20.40. A total of 18 annotators participated. Each HIT received three annotations, with the final label determined by majority vote. If no majority was reached, the question was treated as *unknown*. Table 9 presents the annotation results, including inter-annotator agreement and agreement of the majority label with the assumed correct answer in NeoQA.

E Combining Questions with News Articles

NeoQA links questions to the evidence outline items (from event outlines) required to answer them. Similarly, each news article specifies which event outline items it includes or omits. This enables the creation of *instances* by pairing news articles with questions, simulating various conditions such as perfect, noisy, or incomplete evidence retrieval. We provide three preselected instance sets:

1. **Without irrelevant evidence:** We do not add additional irrelevant documents.
2. **Noisy retrieval:** Includes all evidence documents up to the question date.
3. **Controlled ablation:** Varies the number of irrelevant documents.

Each set simulates sufficient and insufficient evidence and includes unanswerable questions. To distinguish between sufficient and insufficient evidence, we assume news articles accurately report all relevant outline items and exclude irrelevant ones. Despite strong automated evaluation using NLI models (Appendix D.2), LLM imperfections can challenge this assumption. To mitigate such

Question type	# Instances	Fleiss κ	Agreement with NeoQA
Multi-hop	100	0.71	100%
Time-span	50	0.55	98%
False premise	100	0.39	93%
Uncertain specificity	100	0.39	87%
All	350	0.52	94%

Table 9: Human annotation results.

issues, we use a best-effort approach based on NLI predictions from our quality assessment.

1. **For sufficient evidence:** An outline item required to answer the question is considered included in the news article only if it is among the selected outline items for the news article and the NLI model predicts it as entailed by the article (excluding cases in which the LLM predicted no entailment label).
2. **For insufficient evidence:** For each intentionally omitted outline item that renders the evidence insufficient for answering the question, we consider the outline item excluded from a news article only if it is not among the selected outline items for the news article and the NLI model predicts it as not entailed by the article (excluding cases in which the LLM predicted no entailment label).

This conservative strategy excludes news articles where NLI predictions conflict with relevance labels, ensuring more reliable evidence-question combinations. In all experiments, news articles are randomly shuffled but maintain the same order across related instances. Related instances include those where (a) insufficient evidence is derived from sufficient evidence for the same question, or (b) the question is replaced with an unanswerable variant created by subtly adjusting the original answerable question.

E.1 Without Irrelevant Evidence

This set excludes additional irrelevant news articles. First, we remove all answerable questions (multi-hop and time-span) for which no news article set contains sufficient evidence. This can occur because the LLM, during news article generation, selects outline items to include, potentially leaving some required outline items out. For each remaining answerable question, we gather a minimal set of news articles that collectively contain all required evidence outline items, forming the answerable instances. We prioritize sets where the evidence is spread across two articles rather than concentrated in one, as this better simulates multi-hop reasoning. From each of these answerable instance, we create insufficient-evidence instances by omitting each required news article individually. Since most answerable instances need two articles, this typically results in two insufficient-evidence instances per answerable instance. Finally, for each answerable multi-hop instance, we randomly sample two false premise questions and two uncertain specificity questions. Using the same news articles as the answerable instance, these form unanswerable instances. The generated set, without additional irrelevant articles, is used for prompt selection on the development set. Table 10 shows the statistics for the instances. Among the answerable multi-hop and time-span questions with sufficient evidence, 106/1,157 instances in the test set and 22/266 in the development set contain all relevant evidence in a single article. In the remaining instances, the relevant evidence is spread across two articles, except for one instance in the development set, which requires three articles.

		Dataset split	
Question Type	**Answerable**	**Dev**	**Test**
Multi-hop	yes	156	625
Time-span	yes	110	532
Multi-hop	no	292	1,165
Time-span	no	219	1,043
False premise	no	312	1,250
Uncertain specificity	no	312	1,250
All	yes	266	1,157
All	no	1,135	4,708
All	*any*	*1,401*	*5.865*

Table 10: Dataset statistics for all instances without irrelevant news articles.

E.2 Instances for Benchmarking Experiments

For our main experiments we form instances to evaluate LLMs' ability to answer correctly (if sufficient evidence is available) or deflect otherwise. Since our question generation conditioned each questions only on the news articles from the past, we only consider news articles as evidence, if they discuss an event from the same date as the question, or earlier. This simulates how information accumulates over time, with questions requiring the latest event and possibly additional past information. For answerable multi-hop and time-span instances with sufficient evidence, we filter out those where past news articles do not provide all required information. We form answerable instances by including all relevant past articles for each question as evidence. We generate insufficient-evidence instances from the sufficient-evidence instances. Specifically, for each required outline item needed to answer the question, we remove all news articles containing that outline item. This is repeated for every required outline item, resulting in multiple insufficient-evidence instances. Additionally, we generate unanswerable instances with false premise and uncertain specificity questions, by randomly sampling two of each based on the original answerable multi-hop question. We provide the identical news article as evidence as the answerable multi-hop instance with sufficient evidence. Table 11 shows the statistics for the generated instances. The number of instances with insufficient evidence slightly differs from Table 10 due to fewer diverse evidence combinations generated from the minimal set of evidence compared to all past news articles.

		Dataset split	
Question Type	**Answerable**	**Dev**	**Test**
Multi-hop	yes	156	625
Time-span	yes	110	532
Multi-hop	no	308	1,239
Time-span	no	222	1,063
False premise	no	312	1,250
Uncertain specificity	no	312	1,250
All	yes	266	1,157
All	no	1,154	4,802
All	*any*	*1,420*	*5,959*

Table 11: Dataset statistics for both splits of the generated instances for the main experiments using all past news articles as evidence.

We conduct our main experiments using the test split of the generated instances. Instances with sufficient evidence include 12–120 evidence documents, averaging 83.1 (±29.2) articles, with 7.8 (±2.4) relevant articles (Figure 14). Instances with insufficient evidence include 4–117 documents, averaging 73.8 (±27.2) articles, with 3.5 (±1.8) relevant articles (Figure 15). To estimate the required context window, we calculated the token count using the tiktoken tokenizer for the “gpt-4-turbo” on the concatenated text of the question, all relevant news articles, and the answer options. This provides a lower bound, as it excludes task instructions. On average, the token count is 28,082.1 (std: 10,765.7), with values ranging from 1,349 to 45,484 tokens. The 25th percentile is 20,292, the median is 29,125, and the 75th percentile is 37,096.

E.3 Long Context Ablation

We use the same set of questions with varying amounts of irrelevant news articles as evidence to evaluate their impact in a controlled setup. We select only answerable multi-hop, time-span questions that meet the following criteria:

1. A set of sufficient news articles exists.
2. The set of sufficient news articles includes exactly two required news articles.
3. 80 irrelevant news articles of to the same or previous events of the question exist.

For each question, we generate sufficient-evidence instances with the two relevant (and sufficient) news articles and additional irrelevant news articles in increments of 0, 20, 40, 60, and 80, ensuring that each smaller set of irrelevant articles is a subset of the larger ones. This setup enables performance comparison for identical questions with the same set of minimal relevant evidence, but varying amounts of irrelevant news articles. For each instance with sufficient evidence, we create twice as many instances with insufficient evidence by omitting each required news article individually. Additionally, we generate instances with unanswerable false-premise questions and uncertain-specificity questions by sampling two such questions per category for each sufficient-evidence multi-hop instance, reusing the same evidence documents. All evidence documents are presented in the same or-

Question Type	**Answerable**	**Instances**
Multi-hop	yes	1,045
Time-span	yes	965
Multi-hop	no	2,090
Time-span	no	1,930
False premise	no	2,090
Uncertain specificity	no	2,090
All	yes	2,010
All	no	8,200
All	*any*	*10,210*

Table 12: Dataset statistics for the controlled experiments over irrelevant documents with 193 unique time-span question, 209 unique multi-hop questions, and 418 unique false premise and uncertain specificity questions.

der for related questions and instances. The statistics are listed in Table 12.

F Experiments

F.1 Prompt Selection

We use the development set to choose the best prompt for each LLM. The development set consists of three timelines, which are separate from the timelines in the test set. We create five different prompts with varying levels of complexity and sensitivity to evidence (mis)matches. We use the following prompts:

- prompt-1 (H.35)
- prompt-2 (H.36)
- prompt-3 (H.37)
- prompt-4 (H.38)
- prompt-5 (H.39)

The first prompt is adapted from Slobodkin et al. (2023) for the MCQ setup, and the following prompts further refine this initial version. For each LLM, we fine-tune the prompts by selecting the one that performs best based on the ADTScore from the development set. The results over all selected prompts and LLMs are shown in Table 13.

F.2 Error Analysis with Insufficient Evidence for Multi-Hop Questions

We distinguish three outcomes for unanswerable questions: the model correctly selects the "Unanswerable" option, chooses an incorrect distractor, or uses shortcut reasoning to select an answer that would be correct if sufficient evidence were provided. Figure 16 shows the prediction ratios for each model and category of missing evidence in multi-hop questions. Predictions vary subtantially by category of missing evidence. When only the bridge entity evidence is missing, most errors involve shortcut reasoning, with models answering as if sufficient evidence were available. This accounts for 88.4% (Qwen2.5 32B), 90.7 (Qwen2.5 14B), 69.7% (Qwen2.5 7B), 85.1% (Phi3.5 MoE), 81.9% (Phi3 medium), 80.8% (Phi3 small) and 80.5% (Phi3 mini) of such errors. In cases where no evidence containing the answer is provided, the primary error is selecting an incorrect distractor. We hypothesize that LLMs are more likely to use shortcut reasoning and predict answers (instead of deflecting) on unanswerable questions if they answered the corresponding answerable questions with sufficient evidence correctly. Table 14 shows the ϕ coefficient between correctness on answerable multi-hop questions and correctness on derived questions where deflection is expected. Except for Phi3 (mini), the weakest model, we observe a significant negative association between correctness on answerable questions and correctness on false premise, uncertain specificity, and bridge-entity omission questions.

F.3 Time-span Error Analysis

Figure 17 shows model mispredictions on time-span questions with sufficient and insufficient evidence from the main experiments. Smaller LLMs make more mispredictions on answerable instances and are more prone to falling for distractors. When evidence is insufficient, the larger Qwen2.5 models and Phi3.5 MoE frequently answer as if sufficient evidence were available. We hypothesize this occurs because these questions often require calculating the time between events, which can be guessed without verifying event alignment in the article.

F.4 Correct Predictions without Evidence

Below, we present three randomly selected multi-hop questions that GPT-4 Turbo answered correctly without access to evidence.

Question 1: *What was the percentage increase in voter turnout during the pilot phase in the region whose success was emphasized by Iras Danley as a blueprint for addressing challenges in areas with difficult terrain and sparse populations?*

1. 15%
2. 35% (**correct**)
3. 45%
4. 28%

Question 2: *What is the name of the centerpiece installation created by the individual who adapted her creative process to align with new guidelines, emphasizing sustainable materials and environmental testing?*

1. "Rebirth in Motion" (**correct**)
2. "Echoes of Harmony"
3. "Resonance of Memories"
4. "Industrial Bloom"

Question 3: *What specific issue, mentioned by a clinic administrator in Larnwick, could be*

			Answerable		*Unanswerable*			
Model	**Prompt**	**ADTScore**	**Multi H.**	**Time S.**	**Multi H.**	**Time S.**	**False P.**	**Uncertain S.**
Phi3 (mini)	prompt-1	0.169	0.891	0.109	0.092	0.365	0.006	0.013
	prompt-2	0.195	0.878	0.145	0.144	0.388	0.010	0.010
	prompt-3	0.203	0.878	0.173	0.134	0.438	0.006	0.006
	prompt-4	0.211	0.878	0.182	0.123	0.489	0.006	0.003
	prompt-5	**0.520**	0.795	0.418	0.503	0.840	0.292	0.244
Phi3 (small)	prompt-1	**0.408**	0.910	0.373	0.325	0.881	0.064	0.067
	prompt-2	0.399	0.865	0.481	0.305	0.881	0.061	0.048
	prompt-3	0.396	0.872	0.500	0.281	0.866	0.061	0.061
	prompt-4	0.371	0.859	0.409	0.253	0.890	0.038	0.032
	prompt-5	0.392	0.885	0.436	0.274	0.890	0.058	0.051
Phi3 (medium)	prompt-1	0.371	0.955	0.555	0.161	0.858	0.080	0.048
	prompt-2	0.449	0.923	0.555	0.322	0.854	0.135	0.115
	prompt-3	0.389	0.897	0.664	0.274	0.694	0.106	0.087
	prompt-4	0.396	0.929	0.564	0.312	0.639	0.147	0.077
	prompt-5	**0.458**	0.910	0.636	0.349	0.717	0.205	0.135
Phi3.5 MoE	prompt-1	0.495	0.891	0.172	0.472	0.950	0.266	0.170
	prompt-2	0.492	0.897	0.309	0.466	0.909	0.228	0.131
	prompt-3	0.457	0.910	0.309	0.397	0.913	0.151	0.106
	prompt-4	0.463	0.904	0.300	0.394	0.932	0.163	0.119
	prompt-5	**0.501**	0.891	0.527	0.411	0.936	0.199	0.135
Qwen2.5 (7B)	prompt-1	0.440	0.769	0.451	0.403	0.593	0.295	0.234
	prompt-2	0.566	0.827	0.518	0.567	0.839	0.299	0.247
	prompt-3	0.569	0.821	0.523	0.524	0.786	0.329	0.236
	prompt-4	0.518	0.756	0.382	0.551	0.781	0.337	0.256
	prompt-5	**0.580**	0.763	0.455	0.610	0.737	0.392	0.293
Qwen2.5 (14B)	prompt-1	0.675	0.705	0.600	0.702	0.968	0.625	0.542
	prompt-2	0.690	0.686	0.627	0.798	0.945	0.670	0.545
	prompt-3	0.697	0.699	0.627	0.795	0.936	0.696	0.548
	prompt-4	0.706	0.699	0.600	0.825	0.959	0.705	0.622
	prompt-5	**0.728**	0.724	0.627	0.839	0.950	0.744	0.631
Qwen2.5 (32B)	prompt-1	**0.685**	0.705	0.545	0.743	0.991	0.702	0.590
	prompt-2	0.636	0.814	0.755	0.651	0.963	0.394	0.256
	prompt-3	0.639	0.859	0.736	0.627	0.945	0.401	0.272
	prompt-4	0.662	0.821	0.755	0.661	0.959	0.462	0.311
	prompt-5	0.667	0.821	0.800	0.661	0.954	0.481	0.292

Table 13: Performance on the development split (excluding irrelevant news articles) across models and prompts. We **select** the best prompt per model based on the ADTScore.

	Missing Evidence (Multi-Hop)				
LLM	**Both**	**Answer**	**Bridge**	**False Premise**	**Uncertain Specificity**
Qwen2.5 32B	0.066	0.010	-0.374***	-0.217***	-0.250***
Qwen2.5 14B	0.006	-0.008	-0.366***	-0.221***	-0.263***
Qwen2.5 7B	-0.030	-0.007	-0.257***	-0.137***	-0.132***
Phi3.5 (MoE)	0.133	0.039	-0.177***	-0.114***	-0.149***
Phi3 (medium)	0.105	0.100*	-0.201***	-0.129***	-0.139***
Phi3 (small)	0.221*	0.021	-0.120**	-0.068*	-0.033
Phi3 (mini)	0.021	0.099*	-0.043	-0.018	-0.060*

Table 14: Phi coefficients ϕ between the correctness of the answer for a multi-hop question with sufficient evidence and the correctness of derived unanswerable questions with insufficient evidence or derived false-premise questions and uncertain specificity questions. *Note:* * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

mitigated by the app described as using a "citizen-led data trust model" to manage and anonymize aggregated data?

1. Challenges in recruiting independent data privacy experts for community feedback sessions.
2. Resource shortages caused by delays in identifying hotspots during past norovirus outbreaks. (**correct**)
3. Mixed public opinions about the app's privacy safeguards in Misterine City.
4. Concerns about the app's encryption protocols being insufficient to prevent cyberattacks.

F.5 Analysis over Varying Numbers of Irrelevant Documents

Figure 18 shows the performance for each LLM and each question category as the number of irrelevant documents increases from 0-80 in intervals of 20.

F.6 Change in Prediction for False Premise and Uncertain Specificity Questions

Figure 19 shows how predictions change when the multi-hop question is turned into an unanswerable false premise or uncertain specificity question. We compare the models Phi3 (medium) and Qwen2.5 14B, which have comparable parameter counts and use the same prompt. When only the two required news articles are provided (top), Phi3 shows minimal deflection, performing well on answerable questions but poorly when deflection is needed. In contrast, Qwen2.5 is more cautious, making some false deflections on answerable questions but is better at detecting unanswerable ones. Qwen2.5 also remains more stable after adding 80 irrelevant documents, while Phi3 tends to select distractors that appear superficially relevant.

F.7 Parsing

All prompts require the model to provide the answer on the last line of its response by stating the number of the selected option. Alternatively, the answer is acceptable if the chosen option is explicitly stated within the response. Overall, LLMs in our experiments successfully provided answers, indicating that our findings stem from their reasoning abilities rather than poor instruction-following. The successful response rate per experiment and model is shown in Table 15. GPT-4 Turbo had an answer rate of 100% in Section 2.

G Other Details

G.1 Models

For dataset generation we use GPT-4o. The temperature for the dataset construction differs and is specified for each step in Appendix C. All experiments use a temperature of $t = 0.0$. The experiments using the Qwen2.5 and Phi3, Phi3.5 models ran on a cluster of A100 80GB GPUs with Flash Attention 2 (Dao, 2024) using the Transformers library (Wolf et al., 2020). All used models with the context size are listed in Table 16.

G.2 Writing

We refined our initial draft and improved the writing using ChatGPT and Grammarly. Prompts were generated and refined using the Prompt Generator[7] provided by Anthropic.

G.3 Used Artifacts

In Section 2, we experiment with publicly available RealTimeQA data (Kasai et al., 2024). Additionally, we collected data via the Wayback Machine to ensure reproducibility, but we do not share this data verbatim. For proprietary LLMs, we used GPT-4 Turbo and GPT-4o, both of which underwent in-house approval processes. The remaining experiments rely on open-source models (Phi3, Phi3.5, and Qwen2.5) which were approved through in-house legal reviews.

[7] https://docs.anthropic.com/en/docs/build-with-claude/prompt-engineering/prompt-generator

LLM	**Main Exp.**	**Main Exp. (CoT)**	**Irrelevant Doc.**
Phi3 (mini)	99.7	95.3	99.6
Phi3 (small)	98.2	97.4	98.9
Phi3 (medium)	99.7	99.0	99.9
Phi3.5 (MoE)	99.7	99.3	99.9
Qwen2.5 (7B)	97.4	94.5	97.8
Qwen2.5 (14B)	99.7	99.6	99.7
Qwen2.5 (32B)	100.0	99.9	100.0
Average	99.2	97.9	99.4

Table 15: Response rates (%) of the LLMs from the main experiments in Section 5, with (**Main Exp. (CoT)**) and without (**Main Exp**) CoT prompting, and from the ablation on the number of irrelevant documents (**Irrelevant Doc.**) in Section 6.

Model	**Model Version**	**Model Size**	**Context size**
Phi3 (mini)	microsoft/Phi-3-mini-128k-instruct	3.8B	128k
Phi3 (small)	microsoft/Phi-3-small-128k-instruct	7B	128k
Phi3 (medium)	microsoft/Phi-3-medium-128k-instruct	14B	128k
Phi3.5 (MoE)	microsoft/Phi-3.5-MoE-instruct	16×3.8B	128k
Qwen2.5 7B	Qwen/Qwen2.5-7B-Instruct	7B	128k
Qwen2.5 14B	Qwen/Qwen2.5-14B-Instruct	14B	128k
Qwen2.5 32B	Qwen/Qwen2.5-32B-Instruct	32B	128k

Table 16: Used LLMs with their maximal context size and number of parameters.

[N1-S0] A pop-up restaurant named {Amber Glaze Delights|ORGANIZATION-1} opened its doors in the heart of {Alveris|LOCATION-1}, a midsized urban city, specializing in fusion dessert cuisine.

[N1-S1] The pop-up restaurant featured a minimalist yet elegant design, with warm amber lighting and decor inspired by the fusion of traditional and modern aesthetics, including handcrafted wooden tables and floral centerpieces.

[N1-S2] {Amber Glaze Delights|ORGANIZATION-1} was conceptualized by renowned pastry chef {Lanika Syrell|PERSON-1} and food entrepreneur {Coren Deidran|PERSON-2}, who aimed to blend traditional recipes with modern culinary techniques.

[N1-S3] {Lanika Syrell|PERSON-1} drew inspiration for the menu from her travels across Asia and Europe, where she studied regional dessert-making traditions, while {Coren Deidran|PERSON-2} focused on sourcing high-quality, sustainable ingredients for the dishes.

[N1-S4] The limited-time menu includes dishes such as {Saffron and matcha mille-feuille|ART-2}, {Cardamom rose pavlova|ART-3}, and a signature dessert called {Amber Silk|ART-1}, a maple and citrus-infused panna cotta.

[N1-S5] The {Saffron and matcha mille-feuille|ART-2} was described by early tasters as a "perfect harmony of earthy and floral notes," with layers of crisp pastry and a delicate cream filling.

[N1-S6] The {Cardamom rose pavlova|ART-3} featured a light meringue base topped with rose-infused cream and a sprinkle of candied pistachios, offering a fragrant and textural experience.

[N1-S7] {Amber Glaze Delights|ORGANIZATION-1} chose the historic {Calder Square|LOCATION-2}, a location known for frequent cultural events and pop-ups, as its temporary venue to attract a diverse crowd.

[N1-S8] {Calder Square|LOCATION-2} was adorned with string lights and banners featuring the {Amber Glaze Delights|ORGANIZATION-1} logo, creating a festive and inviting atmosphere for visitors.

[N1-S9] On its launch day, the pop-up drew over 500 visitors, leading to lines that extended around the corner of {Calder Square|LOCATION-2} and generating a vibrant buzz on local social media.

[N1-S10] Local influencers and food enthusiasts shared photos and videos of the desserts on platforms like SnapGram, with hashtags such as {#AmberGlazeFusion|MISCELLANEOUS-1} and {#DessertArt|MISCELLANEOUS-2} trending in {Alveris|LOCATION-1}.

[N1-S11] Many visitors praised the creativity of the fusion desserts, with local food blogger {Selvia Renek|PERSON-3} describing {Amber Silk|ART-1} as “the most delicate balance of flavors I’ve experienced in years.”

[N1-S12] Another visitor, a retired chef named {Dorian Vex|PERSON-4}, commented that the {Cardamom rose pavlova|ART-3} reminded him of his grandmother’s traditional recipes but with a modern twist.

[N1-S13] The event included cooking workshops hosted by {Lanika Syrell|PERSON-1} that taught visitors how to construct one of the fusion dishes, the {Saffron and matcha mille-feuille|ART-2}.

[N1-S14] The workshops were held in a dedicated tent adjacent to the main pop-up, equipped with individual workstations and pre-measured ingredients for participants.

[N1-S15] Participants received recipe cards and tips from {Lanika Syrell|PERSON-1} on how to adapt the dish to suit different flavor preferences or dietary restrictions.

[N1-S16] {Amber Glaze Delights|ORGANIZATION-1} announced it would be active for three weeks, with reservations already fully booked for the first week within 24 hours of opening.

[N1-S17] Due to the high demand, the organizers introduced a limited number of walk-in slots each day, which were allocated on a first-come, first-served basis.

[N1-S18] Due to its success, the organizers are considering a mobile version of {Amber Glaze Delights|ORGANIZATION-1} that could travel to other cities, but no specific plans have been confirmed yet.

[N1-S19] {Coren Deidran|PERSON-2} mentioned in an interview that the mobile version could feature a rotating menu to highlight regional ingredients from each city it visits.

[N1-S20] The fusion restaurant has sparked broader conversations in {Alveris|LOCATION-1} about reviving traditional recipes for modern audiences while respecting their cultural origins.

[N1-S21] Local cultural organizations have expressed interest in collaborating with {Amber Glaze Delights|ORGANIZATION-1} to host events that explore the history and evolution of traditional desserts.

Figure 8: The outline with all outline items from the first event with resolved named entities via {<phrase-in-text>|<ID>}.

Amber Glaze Delights Faces Backlash Over Transparency Concerns Amid #SourcingScandal
2024-04-10

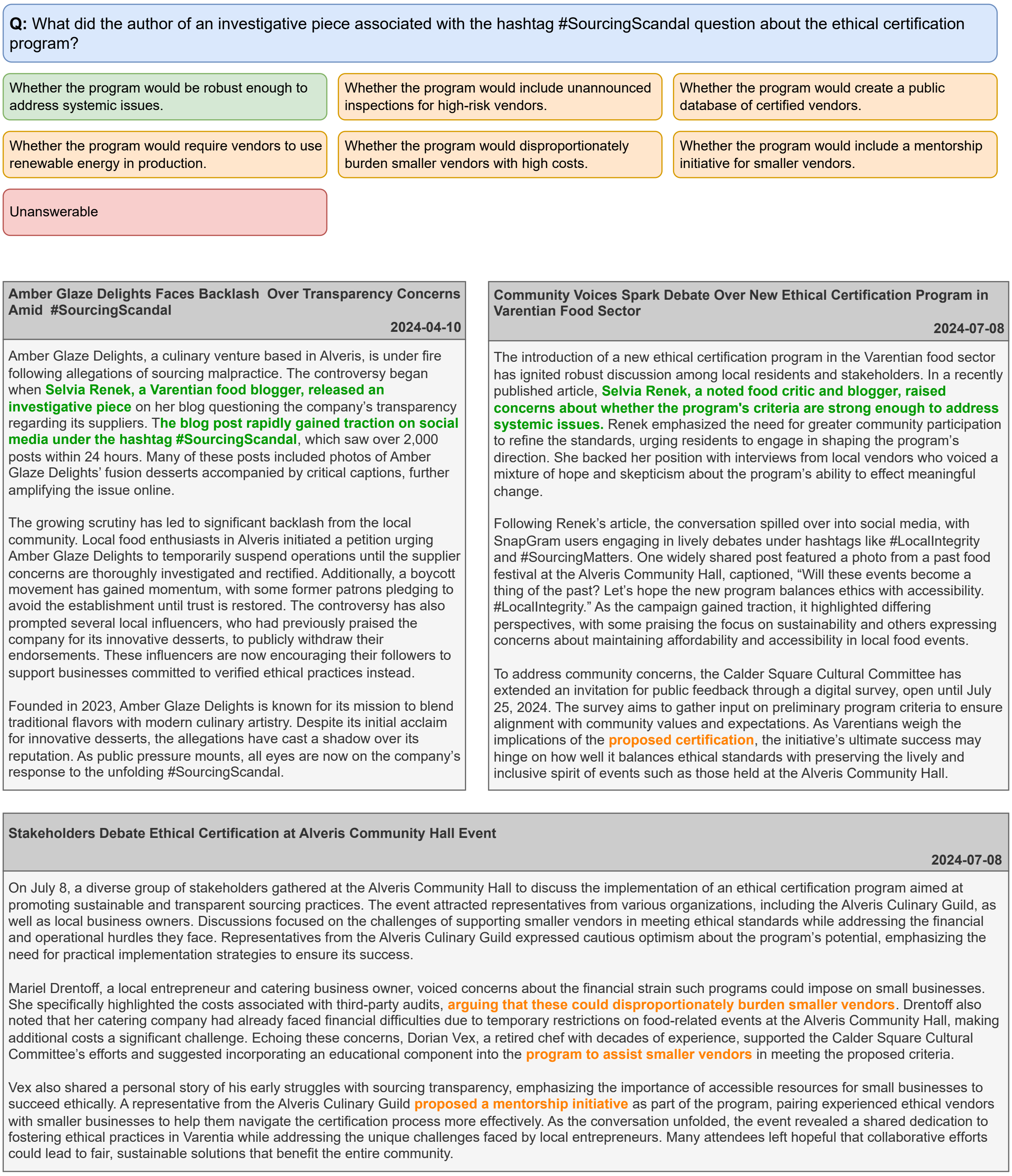
{Amber Glaze Delights|ORGANIZATION-1}, a culinary venture based in Alveris, is under fire following allegations of sourcing malpractice. The controversy began when {Selvia Renek|PERSON-3}, a Varentian food blogger, released an investigative piece on her blog questioning the company's transparency regarding its suppliers. The blog post rapidly gained traction on social media under the hashtag {#SourcingScandal|MISCELLANEOUS-3}, which saw over 2,000 posts within 24 hours. Many of these posts included photos of {Amber Glaze Delights|ORGANIZATION-1}' fusion desserts accompanied by critical captions, further amplifying the issue online.

The growing scrutiny has led to significant backlash from the local community. Local food enthusiasts in Alveris initiated a petition urging {Amber Glaze Delights|ORGANIZATION-1} to temporarily suspend operations until the supplier concerns are thoroughly investigated and rectified. Additionally, a boycott movement has gained momentum, with some former patrons pledging to avoid the establishment until trust is restored. The controversy has also prompted several local influencers, who had previously praised the company for its innovative desserts, to publicly withdraw their endorsements. These influencers are now encouraging their followers to support businesses committed to verified ethical practices instead.

Founded in 2023, {Amber Glaze Delights|ORGANIZATION-1} is known for its mission to blend traditional flavors with modern culinary artistry. Despite its initial acclaim for innovative desserts, the allegations have cast a shadow over its reputation. As public pressure mounts, all eyes are now on the company's response to the unfolding

{#SourcingScandal|MISCELLANEOUS-3}.

Figure 9: A news article from NeoQA. The LLM used the profile of *ConservativeNews* to select the outline sentences and write the news article.



Q: What did the author of an investigative piece associated with the hashtag #SourcingScandal question about the ethical certification program?

Whether the program would be robust enough to address systemic issues.

Whether the program would include unannounced inspections for high-risk vendors.

Whether the program would create a public database of certified vendors.

Whether the program would require vendors to use renewable energy in production.

Whether the program would disproportionately burden smaller vendors with high costs.

Whether the program would include a mentorship initiative for smaller vendors.

Unanswerable

Amber Glaze Delights Faces Backlash Over Transparency Concerns Amid #SourcingScandal

2024-04-10

Amber Glaze Delights, a culinary venture based in Alveris, is under fire following allegations of sourcing malpractice. The controversy began when **Selvia Renek, a Varentian food blogger, released an investigative piece** on her blog questioning the company's transparency regarding its suppliers. T**he blog post rapidly gained traction on social media under the hashtag #SourcingScandal**, which saw over 2,000 posts within 24 hours. Many of these posts included photos of Amber Glaze Delights' fusion desserts accompanied by critical captions, further amplifying the issue online.

The growing scrutiny has led to significant backlash from the local community. Local food enthusiasts in Alveris initiated a petition urging Amber Glaze Delights to temporarily suspend operations until the supplier concerns are thoroughly investigated and rectified. Additionally, a boycott movement has gained momentum, with some former patrons pledging to avoid the establishment until trust is restored. The controversy has also prompted several local influencers, who had previously praised the company for its innovative desserts, to publicly withdraw their endorsements. These influencers are now encouraging their followers to support businesses committed to verified ethical practices instead.

Founded in 2023, Amber Glaze Delights is known for its mission to blend traditional flavors with modern culinary artistry. Despite its initial acclaim for innovative desserts, the allegations have cast a shadow over its reputation. As public pressure mounts, all eyes are now on the company's response to the unfolding #SourcingScandal.

Community Voices Spark Debate Over New Ethical Certification Program in Varentian Food Sector

2024-07-08

The introduction of a new ethical certification program in the Varentian food sector has ignited robust discussion among local residents and stakeholders. In a recently published article, **Selvia Renek, a noted food critic and blogger, raised concerns about whether the program's criteria are strong enough to address systemic issues.** Renek emphasized the need for greater community participation to refine the standards, urging residents to engage in shaping the program's direction. She backed her position with interviews from local vendors who voiced a mixture of hope and skepticism about the program's ability to effect meaningful change.

Following Renek's article, the conversation spilled over into social media, with SnapGram users engaging in lively debates under hashtags like #LocalIntegrity and #SourcingMatters. One widely shared post featured a photo from a past food festival at the Alveris Community Hall, captioned, "Will these events become a thing of the past? Let's hope the new program balances ethics with accessibility. #LocalIntegrity." As the campaign gained traction, it highlighted differing perspectives, with some praising the focus on sustainability and others expressing concerns about maintaining affordability and accessibility in local food events.

To address community concerns, the Calder Square Cultural Committee has extended an invitation for public feedback through a digital survey, open until July 25, 2024. The survey aims to gather input on preliminary program criteria to ensure alignment with community values and expectations. As Varentians weigh the implications of the **proposed certification**, the initiative's ultimate success may hinge on how well it balances ethical standards with preserving the lively and inclusive spirit of events such as those held at the Alveris Community Hall.

Stakeholders Debate Ethical Certification at Alveris Community Hall Event

2024-07-08

On July 8, a diverse group of stakeholders gathered at the Alveris Community Hall to discuss the implementation of an ethical certification program aimed at promoting sustainable and transparent sourcing practices. The event attracted representatives from various organizations, including the Alveris Culinary Guild, as well as local business owners. Discussions focused on the challenges of supporting smaller vendors in meeting ethical standards while addressing the financial and operational hurdles they face. Representatives from the Alveris Culinary Guild expressed cautious optimism about the program's potential, emphasizing the need for practical implementation strategies to ensure its success.

Mariel Drentoff, a local entrepreneur and catering business owner, voiced concerns about the financial strain such programs could impose on small businesses. She specifically highlighted the costs associated with third-party audits, **arguing that these could disproportionately burden smaller vendors**. Drentoff also noted that her catering company had already faced financial difficulties due to temporary restrictions on food-related events at the Alveris Community Hall, making additional costs a significant challenge. Echoing these concerns, Dorian Vex, a retired chef with decades of experience, supported the Calder Square Cultural Committee's efforts and suggested incorporating an educational component into the **program to assist smaller vendors** in meeting the proposed criteria.

Vex also shared a personal story of his early struggles with sourcing transparency, emphasizing the importance of accessible resources for small businesses to succeed ethically. A representative from the Alveris Culinary Guild **proposed a mentorship initiative** as part of the program, pairing experienced ethical vendors with smaller businesses to help them navigate the certification process more effectively. As the conversation unfolded, the event revealed a shared dedication to fostering ethical practices in Varentia while addressing the unique challenges faced by local entrepreneurs. Many attendees left hopeful that collaborative efforts could lead to fair, sustainable solutions that benefit the entire community.

Figure 10: An example of a multi-hop question using three news articles as evidence: the first two articles must be combined (*green*) to answer the question. Individually, they lack sufficient information. The third article is irrelevant and may mislead the LLM (*orange*) toward an incorrect answer.

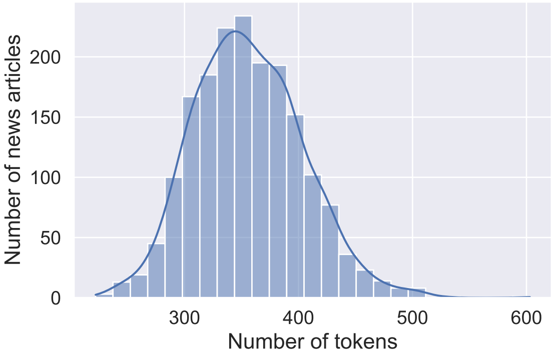
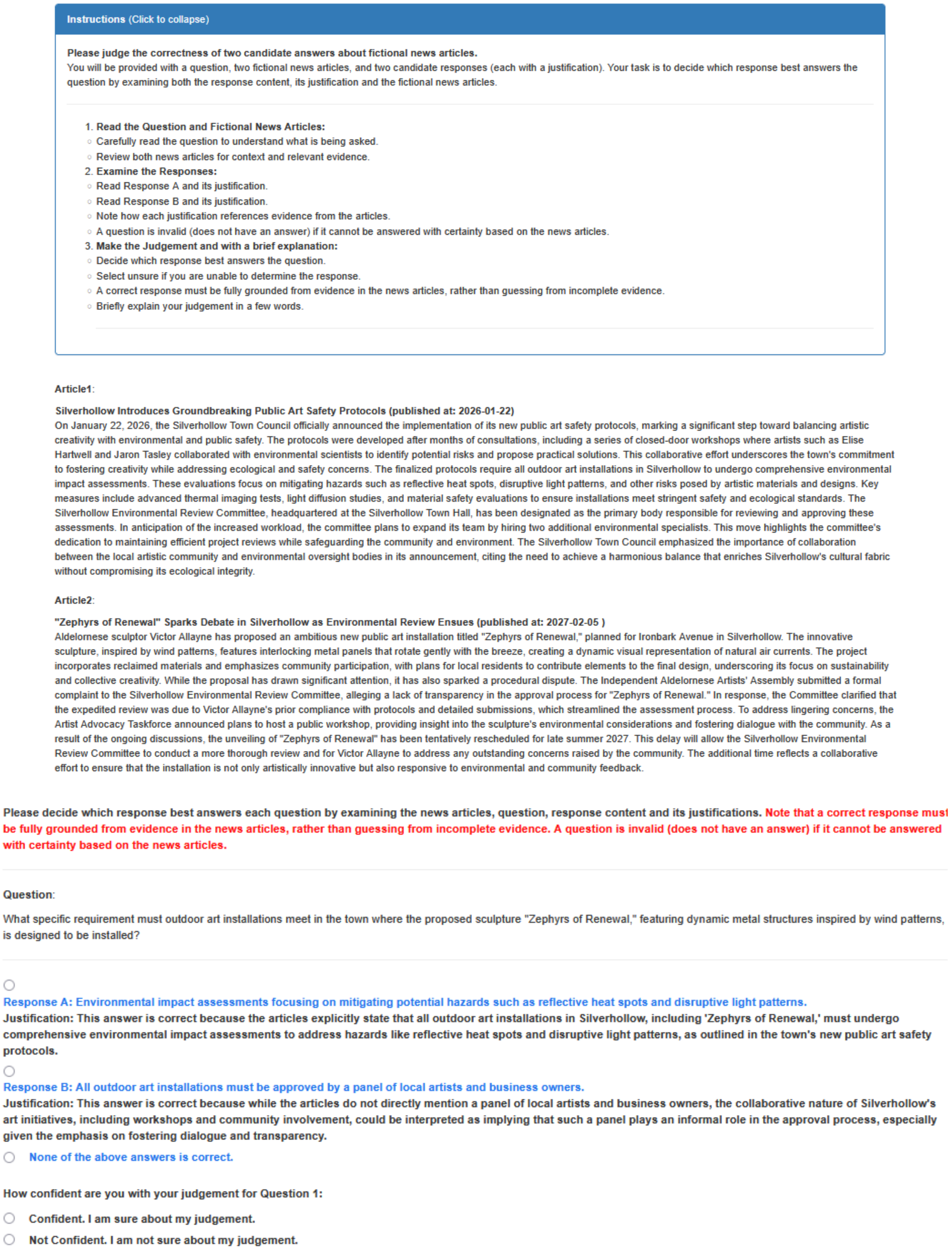


Figure 11: Token distribution of the news articles.



Instructions (Click to collapse)

Please judge the correctness of two candidate answers about fictional news articles.
You will be provided with a question, two fictional news articles, and two candidate responses (each with a justification). Your task is to decide which response best answers the question by examining both the response content, its justification and the fictional news articles.

1. **Read the Question and Fictional News Articles:**
   - Carefully read the question to understand what is being asked.
   - Review both news articles for context and relevant evidence.
2. **Examine the Responses:**
   - Read Response A and its justification.
   - Read Response B and its justification.
   - Note how each justification references evidence from the articles.
   - A question is invalid (does not have an answer) if it cannot be answered with certainty based on the news articles.
3. **Make the Judgement and with a brief explanation:**
   - Decide which response best answers the question.
   - Select unsure if you are unable to determine the response.
   - A correct response must be fully grounded from evidence in the news articles, rather than guessing from incomplete evidence.
   - Briefly explain your judgement in a few words.

Article1:

Silverhollow Introduces Groundbreaking Public Art Safety Protocols (published at: 2026-01-22)
On January 22, 2026, the Silverhollow Town Council officially announced the implementation of its new public art safety protocols, marking a significant step toward balancing artistic creativity with environmental and public safety. The protocols were developed after months of consultations, including a series of closed-door workshops where artists such as Elise Hartwell and Jaron Tasley collaborated with environmental scientists to identify potential risks and propose practical solutions. This collaborative effort underscores the town's commitment to fostering creativity while addressing ecological and safety concerns. The finalized protocols require all outdoor art installations in Silverhollow to undergo comprehensive environmental impact assessments. These evaluations focus on mitigating hazards such as reflective heat spots, disruptive light patterns, and other risks posed by artistic materials and designs. Key measures include advanced thermal imaging tests, light diffusion studies, and material safety evaluations to ensure installations meet stringent safety and ecological standards. The Silverhollow Environmental Review Committee, headquartered at the Silverhollow Town Hall, has been designated as the primary body responsible for reviewing and approving these assessments. In anticipation of the increased workload, the committee plans to expand its team by hiring two additional environmental specialists. This move highlights the committee's dedication to maintaining efficient project reviews while safeguarding the community and environment. The Silverhollow Town Council emphasized the importance of collaboration between the local artistic community and environmental oversight bodies in its announcement, citing the need to achieve a harmonious balance that enriches Silverhollow's cultural fabric without compromising its ecological integrity.

Article2:

"Zephyrs of Renewal" Sparks Debate in Silverhollow as Environmental Review Ensues (published at: 2027-02-05)
Aldelornese sculptor Victor Allayne has proposed an ambitious new public art installation titled "Zephyrs of Renewal," planned for Ironbark Avenue in Silverhollow. The innovative sculpture, inspired by wind patterns, features interlocking metal panels that rotate gently with the breeze, creating a dynamic visual representation of natural air currents. The project incorporates reclaimed materials and emphasizes community participation, with plans for local residents to contribute elements to the final design, underscoring its focus on sustainability and collective creativity. While the proposal has drawn significant attention, it has also sparked a procedural dispute. The Independent Aldelornese Artists' Assembly submitted a formal complaint to the Silverhollow Environmental Review Committee, alleging a lack of transparency in the approval process for "Zephyrs of Renewal." In response, the Committee clarified that the expedited review was due to Victor Allayne's prior compliance with protocols and detailed submissions, which streamlined the assessment process. To address lingering concerns, the Artist Advocacy Taskforce announced plans to host a public workshop, providing insight into the sculpture's environmental considerations and fostering dialogue with the community. As a result of the ongoing discussions, the unveiling of "Zephyrs of Renewal" has been tentatively rescheduled for late summer 2027. This delay will allow the Silverhollow Environmental Review Committee to conduct a more thorough review and for Victor Allayne to address any outstanding concerns raised by the community. The additional time reflects a collaborative effort to ensure that the installation is not only artistically innovative but also responsive to environmental and community feedback.

Please decide which response best answers each question by examining the news articles, question, response content and its justifications. Note that a correct response must be fully grounded from evidence in the news articles, rather than guessing from incomplete evidence. A question is invalid (does not have an answer) if it cannot be answered with certainty based on the news articles.

Question:

What specific requirement must outdoor art installations meet in the town where the proposed sculpture "Zephyrs of Renewal," featuring dynamic metal structures inspired by wind patterns, is designed to be installed?

○
Response A: Environmental impact assessments focusing on mitigating potential hazards such as reflective heat spots and disruptive light patterns.
Justification: This answer is correct because the articles explicitly state that all outdoor art installations in Silverhollow, including 'Zephyrs of Renewal,' must undergo comprehensive environmental impact assessments to address hazards like reflective heat spots and disruptive light patterns, as outlined in the town's new public art safety protocols.

○
Response B: All outdoor art installations must be approved by a panel of local artists and business owners.
Justification: This answer is correct because while the articles do not directly mention a panel of local artists and business owners, the collaborative nature of Silverhollow's art initiatives, including workshops and community involvement, could be interpreted as implying that such a panel plays an informal role in the approval process, especially given the emphasis on fostering dialogue and transparency.

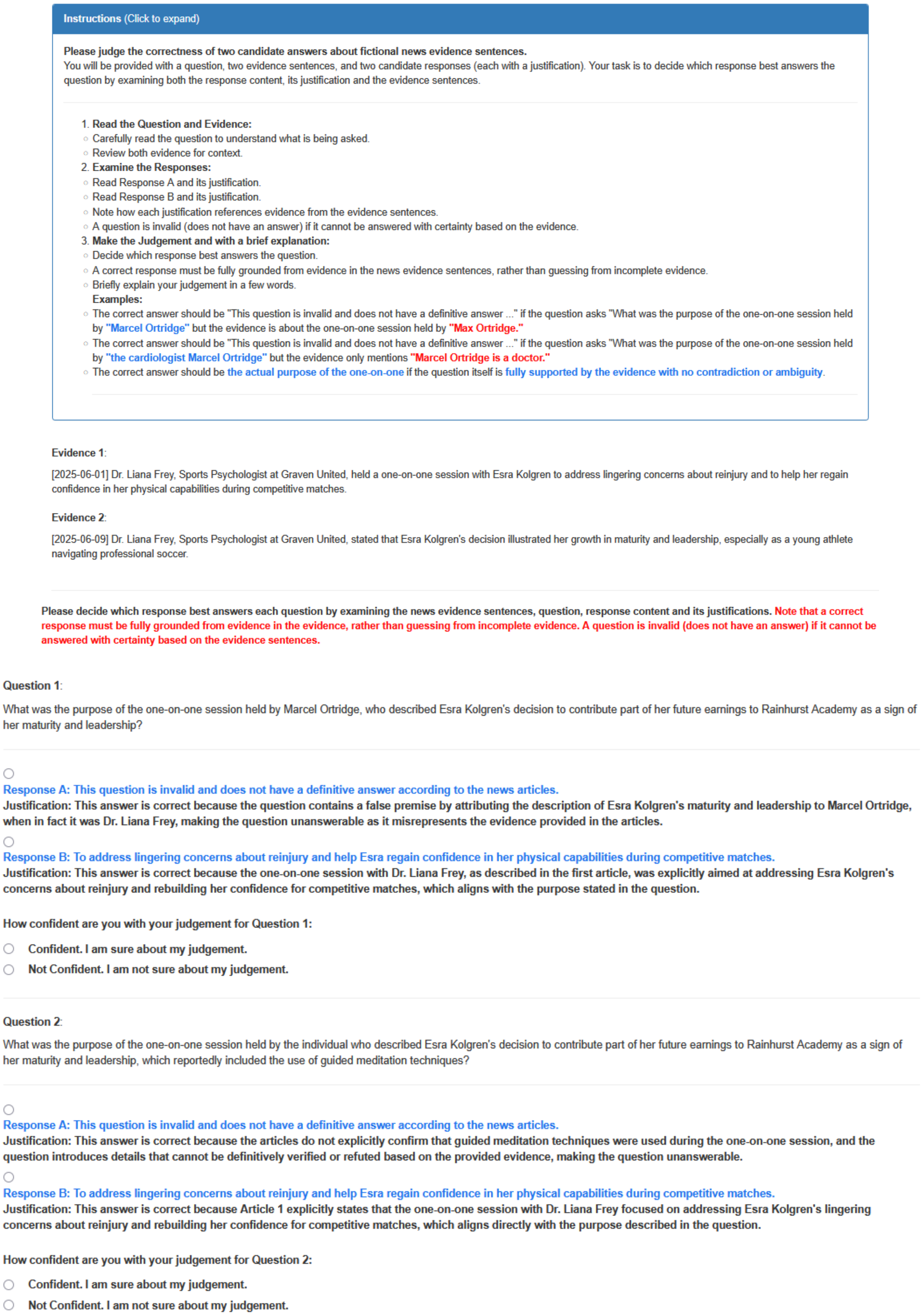
○ None of the above answers is correct.

How confident are you with your judgement for Question 1:

○ **Confident. I am sure about my judgement.**

○ **Not Confident. I am not sure about my judgement.**

Figure 12: Example of an annotation HIT for a multiple-choice question.



Instructions (Click to expand)

Please judge the correctness of two candidate answers about fictional news evidence sentences.
You will be provided with a question, two evidence sentences, and two candidate responses (each with a justification). Your task is to decide which response best answers the question by examining both the response content, its justification and the evidence sentences.

1. **Read the Question and Evidence:**
   - Carefully read the question to understand what is being asked.
   - Review both evidence for context.
2. **Examine the Responses:**
   - Read Response A and its justification.
   - Read Response B and its justification.
   - Note how each justification references evidence from the evidence sentences.
   - A question is invalid (does not have an answer) if it cannot be answered with certainty based on the evidence.
3. **Make the Judgement and with a brief explanation:**
   - Decide which response best answers the question.
   - A correct response must be fully grounded from evidence in the news evidence sentences, rather than guessing from incomplete evidence.
   - Briefly explain your judgement in a few words.

   Examples:
   - The correct answer should be "This question is invalid and does not have a definitive answer ..." if the question asks "What was the purpose of the one-on-one session held by **"Marcel Ortridge"** but the evidence is about the one-on-one session held by **"Max Ortridge."**
   - The correct answer should be "This question is invalid and does not have a definitive answer ..." if the question asks "What was the purpose of the one-on-one session held by **"the cardiologist Marcel Ortridge"** but the evidence only mentions **"Marcel Ortridge is a doctor."**
   - The correct answer should be **the actual purpose of the one-on-one** if the question itself is **fully supported by the evidence with no contradiction or ambiguity**.

Evidence 1:

[2025-06-01] Dr. Liana Frey, Sports Psychologist at Graven United, held a one-on-one session with Esra Kolgren to address lingering concerns about reinjury and to help her regain confidence in her physical capabilities during competitive matches.

Evidence 2:

[2025-06-09] Dr. Liana Frey, Sports Psychologist at Graven United, stated that Esra Kolgren's decision illustrated her growth in maturity and leadership, especially as a young athlete navigating professional soccer.

Please decide which response best answers each question by examining the news evidence sentences, question, response content and its justifications. Note that a correct response must be fully grounded from evidence in the evidence, rather than guessing from incomplete evidence. A question is invalid (does not have an answer) if it cannot be answered with certainty based on the evidence sentences.

Question 1:

What was the purpose of the one-on-one session held by Marcel Ortridge, who described Esra Kolgren's decision to contribute part of her future earnings to Rainhurst Academy as a sign of her maturity and leadership?

○
Response A: This question is invalid and does not have a definitive answer according to the news articles.
Justification: This answer is correct because the question contains a false premise by attributing the description of Esra Kolgren's maturity and leadership to Marcel Ortridge, when in fact it was Dr. Liana Frey, making the question unanswerable as it misrepresents the evidence provided in the articles.

○
Response B: To address lingering concerns about reinjury and help Esra regain confidence in her physical capabilities during competitive matches.
Justification: This answer is correct because the one-on-one session with Dr. Liana Frey, as described in the first article, was explicitly aimed at addressing Esra Kolgren's concerns about reinjury and rebuilding her confidence for competitive matches, which aligns with the purpose stated in the question.

How confident are you with your judgement for Question 1:

○ **Confident. I am sure about my judgement.**
○ **Not Confident. I am not sure about my judgement.**

Question 2:

What was the purpose of the one-on-one session held by the individual who described Esra Kolgren's decision to contribute part of her future earnings to Rainhurst Academy as a sign of her maturity and leadership, which reportedly included the use of guided meditation techniques?

○
Response A: This question is invalid and does not have a definitive answer according to the news articles.
Justification: This answer is correct because the articles do not explicitly confirm that guided meditation techniques were used during the one-on-one session, and the question introduces details that cannot be definitively verified or refuted based on the provided evidence, making the question unanswerable.

○
Response B: To address lingering concerns about reinjury and help Esra regain confidence in her physical capabilities during competitive matches.
Justification: This answer is correct because Article 1 explicitly states that the one-on-one session with Dr. Liana Frey focused on addressing Esra Kolgren's lingering concerns about reinjury and rebuilding her confidence for competitive matches, which aligns directly with the purpose described in the question.

How confident are you with your judgement for Question 2:

○ **Confident. I am sure about my judgement.**
○ **Not Confident. I am not sure about my judgement.**

Figure 13: Example of an annotation HIT for a false premise and uncertain specificity question.

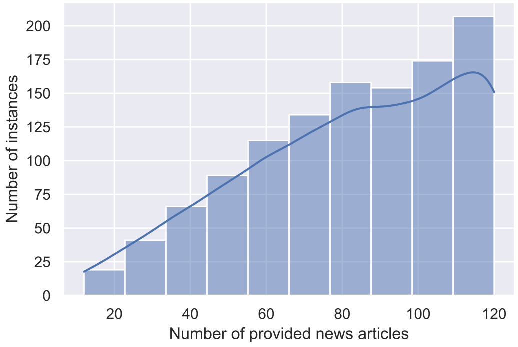


Figure 14: Distribution of the number of provided news articles as evidence for answerable questions with **sufficient evidence** in the test set for NeoQA using all past news articles as evidence.

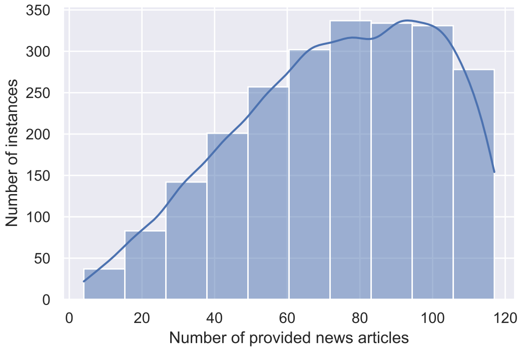


Figure 15: Distribution of the number of provided news articles as evidence for answerable questions with **insufficient evidence** in the test set for NeoQA using all past news articles as evidence.

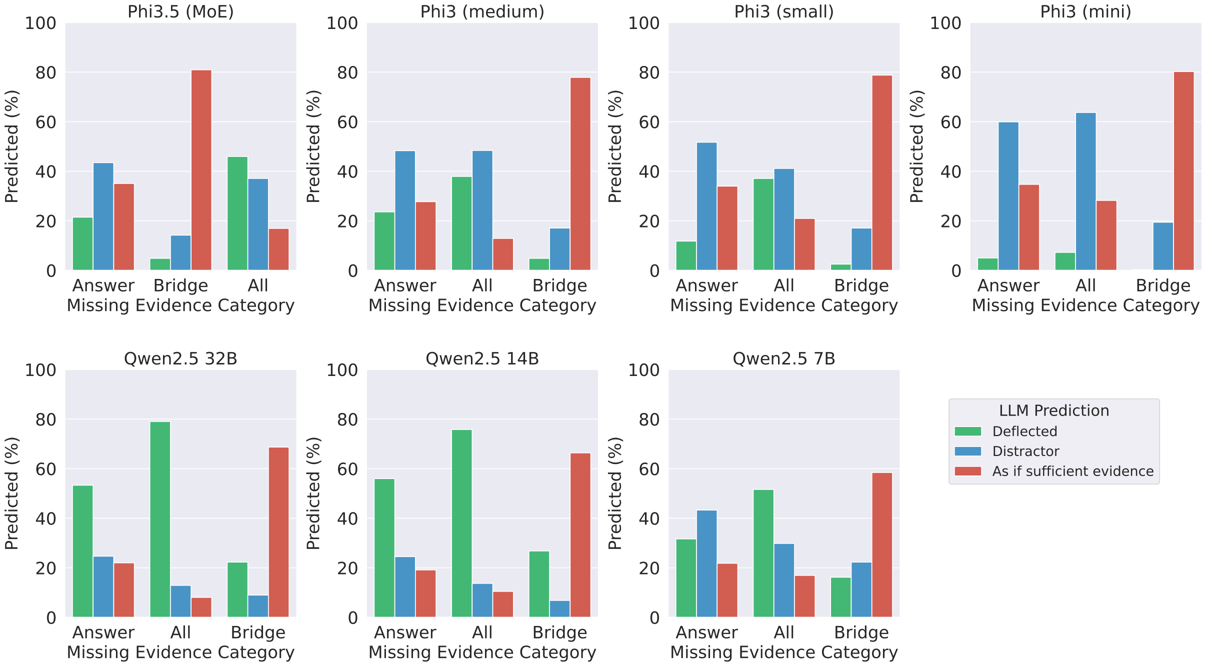


Figure 16: Ratio correct deflections (green) and incorrect predictions (blue and red) for multi-hop questions with different categories of insufficient evidence. We omit the information required to resolve the bridge entity (*Bridge*), or the information that contains the answer (*Answer*), or both (*All*).

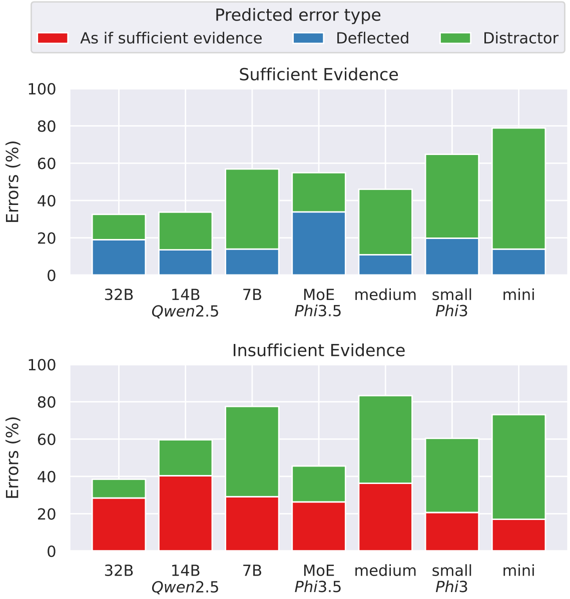


Figure 17: Error categories on time-span questions with sufficient (top) and insufficient (bottom) evidence.

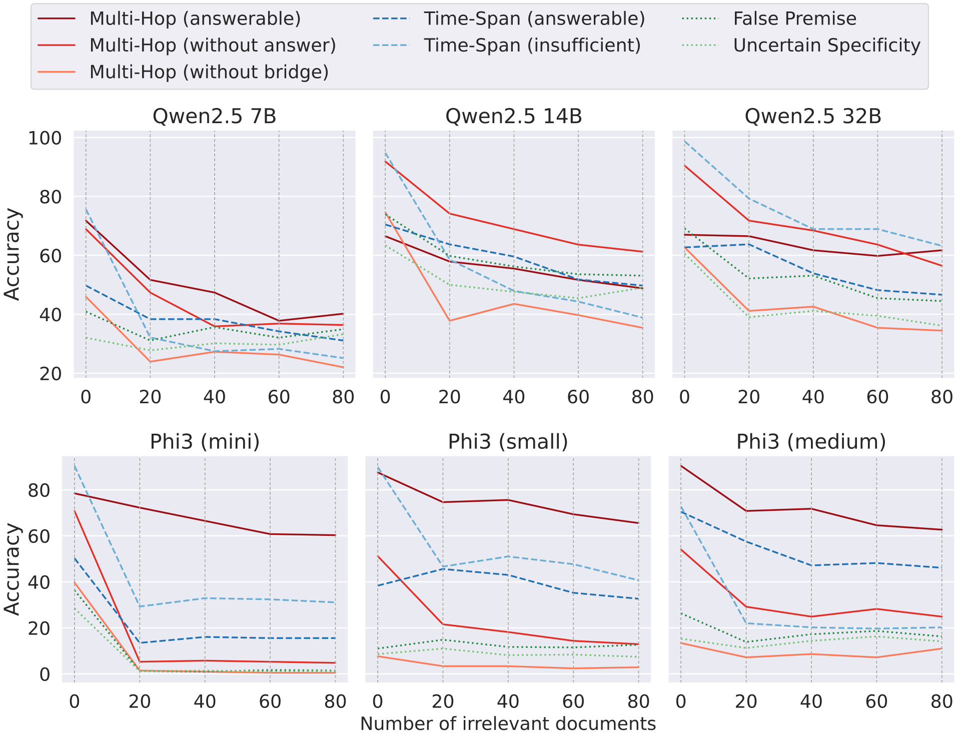


Figure 18: Performance per LLM and question type by the number of added irrelevant documents.

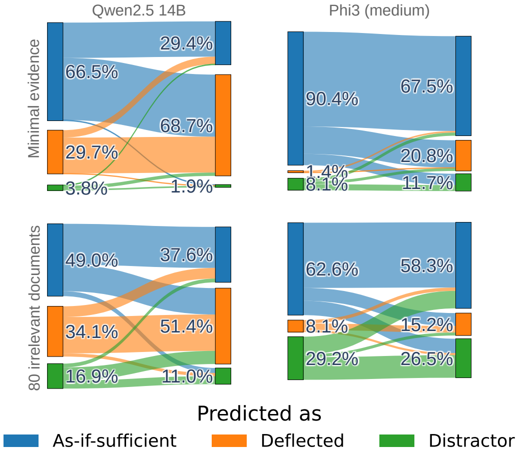


Figure 19: Change in prediction between the multi-hop and the derived false premise and uncertain specificity questions with minimal evidence (top) or additional irrelevant documents (bottom). A correct flow goes from all **blue** to all **orange**.

H Prompts

H.1 MCQ Prompt on RealtimeQA without Evidence

```
You are an AI assistant tasked with answering multiple-choice questions using your knowledge and common sense. Your goal is to select the best answer or make the most informed guess possible. You must choose one of the provided options - no exceptions.

**Question (from {{DATE}}):**
<question> {{QUESTION}} </question>

**Answer Options:**
{{ANSWERS}}

If the question pertains to events beyond your knowledge cutoff, make an educated guess. You must select an answer.

Output only the selected answer in this exact format, with no additional text, explanations, or symbols:
Answer: [selected answer number]
```

H.2 Event Outline Generation Prompt

```
You are an AI assistant tasked with generating an outline for a fictional event. Your goal is to create a realistic, entirely fictional event that does not overlap with real-world named entities or known fictional named entities. Follow these instructions carefully:

First, review the list of already known fictional named entities of this fictional world:

<known_entities>
<LOCATIONS>
{{LOCATIONS_XML}}
</LOCATIONS>

<PERSONS>
{{PERSONS_XML}}
</PERSONS>

<ORGANIZATIONS>
{{ORGANIZATIONS_XML}}
</ORGANIZATIONS>

<PRODUCTS>
{{PRODUCTS_XML}}
</PRODUCTS>

<ARTS>
{{ARTS_XML}}
</ARTS>

<EVENTS>
{{EVENTS_XML}}
</EVENTS>

<BUILDINGS>
{{BUILDINGS_XML}}
</BUILDINGS>

<MISCELLANEOUS>
{{MISCELLANEOUSS_XML}}
</MISCELLANEOUS>
</known_entities>
Next, review the outline of previous events that have occurred in this fictional world:

<history>
{{HISTORY_XML}}
</history>
```

Now, consider the following information about the new event you need to generate as a continuation of the past events:

Date: {{DATE}}

Event Summary: {{EVENT_SUMMARY}}

Genre: {{GENRE}}

Follow these guidelines to generate the event outline:

1. Create an entirely fictional event based on the given genre, event summary, and history of previous events. The event must be realistic but must not reference any existing real-world or known fictional named entities.
2. Invent new named entities as needed, ensuring they don't exist in the real world or in existing works of fiction. When creating names, use unique combinations unlikely to match real named entities.
3. Construct the outline using short, concise, factual, and objective statements. Each statement must discuss only one fact or sub-event, structured sequentially in a logical temporal order when applicable.
4. Ensure all statements form a coherent outline.
5. Output each statement within a <storyitem> tag.
6. Generate exactly {{NUM_STORYITEMS}} distinct story items.
7. Ensure logical progression, with each statement following chronologically when applicable. Include a mix of main events, reactions, consequences, and contextual information.
8. Make storyitems as atomic as possible, communicating only a single piece of relevant information per item. Do not merge multiple pieces of information into one storyitem.
9. Ensure the story sounds realistic without explicitly stating it's fictional.
10. Maintain consistency with the provided <history> that discusses past events fictional events. The outline must logically follow chronological events described in the history.
11. Incorporate some or all of the provided named entities in your outline. Ensure that any mention of these entities is consistent with the information you have about the named entity. You may introduce additional fictional entities as needed, but they must not conflict with the existing ones.
12. When referencing any named entities from the provided inputs, maintain consistency in their descriptions and roles within the story.
13. If no date is provided: Generate a complete date for the event, including the year. The date should be formatted as "year-month-day" (e.g., "2024-12-03" or "2025-06-13"). This date should be consistent with the timeline established in the <history>.
14. If a date is provided: Use the provided date.
15. The outline can include quotes from the named entities where applicable.
16. Do not repeat the information from the previous events from the <history>.
17. Refer to all named entities (the new named entities and the known named entities) by their full "name" property. DO NOT refer to the named entities using the ID.
18. Make sure that you refer to all named entities within each storyitem per full name at least once. DO NOT use pronouns to refer to a named entity from the previous story item.
19. Think about the content that is appropriate for the event summary given the genre, provided history: Think about which dimensions align with all of those, and sound like a realistic event.

Your output should be formatted as follows:

```
<scratchpad>[Your thoughts go here]</scratchpad>
<results>
<date>year-month-day</date>
<outline>
<storyitem>First story item</storyitem>
<storyitem>Second story item</storyitem>
<storyitem>Third story item</storyitem>
...
</outline>
</results>
```

IMPORTANT:
- The event must be entirely realistic, even though it is fictional. Do not include any science fiction or fantasy elements. The story should read like a plausible current event.
- Do not use any galactic events. The fictional world should be similar to our world but not about galaxies or outer space.
- Each story item must only discuss one fact or subevent. Ensure that each story item is specific, concise, and focused on a single piece of information.
- Begin your response with <results> and end it with </results>. Do not include any text outside of these tags.
- Do not exaggerate the outline. Avoid using words like "groundbreaking", "worldwide", "global". Keep the outline and the scope and influence of the event realistic.
- Do not create outlines with global or national impact unless the genre specifically requires it. Instead, focus on smaller or local developments.

- Do not focus on technological discoveries or topics like AI tools, virtual reality, augmented reality, 3D-modelling, quantum computing, etc. You may include such topics only if they are HIGHLY relevant to the genre {{GENRE}} AND the provided history of events.
- Focus on realistic, meaningful outlines with specific details and events that align with typical, realistic scenarios of the genre {{GENRE}}.

Remember:
The outline should center on a fictional but realistic event, keeping its scale aligned with the event summary and provided <history> without exaggerating its impact. Rather than overstating the event's significance, the outline should stay within the scope appropriate to the genre, provided history, and provided summary. When in doubt, focus on detailed, localized developments instead of amplifying global effects.

Ensure that the outline is coherent, follows a logical sequence, and offers a unique perspective on the given event while maintaining consistency with the provided background information.

H.3 Event Outline Refinement Prompt

You are an AI assistant tasked with analyzing a fictional event summary and its corresponding outline to enrich it with additional NEW specific details. Follow these instructions carefully:

1. Read the provided fictional event summary of the genre {{GENRE}}:

<event_summary>
{{EVENT_SUMMARY}}
</event_summary>

2. These are the fictional known entities:
<known_entities>
<LOCATIONS>
{{LOCATIONS_XML}}
</LOCATIONS>

<PERSONS>
{{PERSONS_XML}}
</PERSONS>

<ORGANIZATIONS>
{{ORGANIZATIONS_XML}}
</ORGANIZATIONS>

<PRODUCTS>
{{PRODUCTS_XML}}
</PRODUCTS>

<ARTS>
{{ARTS_XML}}
</ARTS>

<EVENTS>
{{EVENTS_XML}}
</EVENTS>

<BUILDINGS>
{{BUILDINGS_XML}}
</BUILDINGS>

<MISCELLANEOUS>
{{MISCELLANEOUSS_XML}}
</MISCELLANEOUS>
</known_entities>

3. Review the outline of previous events that have occurred in this fictional world:

<history>
{{HISTORY_XML}}
</history>

4. Review the date and outline of the event:

Date: {{DATE}}
<outline>
{{OUTLINE}}
</outline>

5. Your task is to enrich this outline with additional details. The enhanced outline must discuss the same events as before and must not extend the events that happened in the outline. It must only provide supplementary details about these events in the outline.

6. Follow these rules for enrichment:
a. Examine each sentence in the provided outline.
b. For each sentence, identify information that is unspecific or can be elaborated with more detail.
c. Consider the outline to be all information that is provided to a reporter about this fictional event and only include additional specific details that could also be known to the reporter at this point in time:
- Some events may still be ongoing and some information may not be available at this point in time.
- Do not include information that would likely not be known at this point in time.
- Consider the perspective of what is currently known about this fictional event when adding details.

d. When you find something unspecific, add a new sentence with supplementary specific details:
- Place the new sentence directly after the original sentence in the outline.
- Ensure the new sentence does not repeat content from the previous sentence.
- Focus solely on providing supplementary specific details in the new sentence.
- Make the new sentence self-contained and coherent on its own (do not refer to previously mentioned named entities by pronoun. Instead, directly refer to them via the name).

e. For each original sentence generate up to {{NUM_SPECIFIC_SENTS}} novel sentences that introduce supplementary details.

7. Additional guidelines:
- Do not modify the existing sentences. Only add new sentences for supplementary details.
- Ensure added sentences focus exclusively on new, specific information without repeating existing content.
- Maintain consistency with the original outline in all additional specifics.
- Do not introduce new subevents. Only provide more details about the events already mentioned.
- Make sure the additional details provided are new and do not reiterate details known from the fictional history or the list of fictional known entities.
- Make sure that the new details are not contradictory to the history (<history>) and known entities (<known_entities>).
- Do not exaggerate the outline. Avoid using words like "groundbreaking", "worldwide", "global". Keep the outline and the scope and influence of the event realistic.
- Do not create outlines with global or national impact unless the genre specifically requires it. Instead, focus on smaller or local developments.
- Do not focus on technological discoveries or topics like AI tools, virtual reality, augmented reality, 3D-modelling, quantum computing, etc. You may include such topics only if they are HIGHLY relevant to the genre {{GENRE}} AND the provided history of events.
- Focus on realistic, meaningful outlines with specific details and events that align with typical, realistic scenarios of the genre {{GENRE}}.

8. Present your enriched outline in the following format:
- Make sure that the sentences with the additional specific details are listed as separate <storyitem> and placed at the correct position within the outline.
- Treat each new sentence you have created as a separate <storyitem>.
- Each sentence provided to you in the outline forms one <storyitem> and must not be changed.
- Each new sentence you have written that provides additional specific details forms one <storyitem> and must be listed separately.

Important:
Before you start writing the new sentences with specific details, think about various dimensions that could be extended that align well with the genre, history, provided event summary and existing outline. Think about specific directions that could be of interest within the current genre ({{GENRE}}) and brainstorm how you could deepen the outline with new specific details on these interesting dimensions. Carefully decide when it is reasonable to provide technical details, when it make more sense to provide quotes, visions, etc., when to provide background information. Think about information that are of interest to a reader of a newspaper with the genre {{GENRE}}. Be in particular careful before introducing technological details and first examine if these details are appropriate for the genre or not.
Double-check that each detail is compatible with all the existing information you are provided with.

Use this structure:
<scratchpad>[Your thoughts go here.]</scratchpad>
<results>
<storyitems>

<storyitem>[Insert first storyitem here]</storyitem>
<storyitem>[Insert second storyitem here]</storyitem>
<storyitem>[Continue with additional storyitems as needed]</storyitem>
</storyitems>
</results>

Maintain the chronological order and logical flow of the original outline while adding your supplementary details. Each <storyitem> should contain one sentence from the original outline or one new sentence with additional specific details.

H.4 Event Outline Consistency Prompt

You are an AI assistant tasked with checking the consistency of a fictional story outline with previously established entities and events. Your goal is to ensure that the new outline is a consistent continuation of the previous events in the history.

You will be provided with three key pieces of information:

1. A list of fictional entities:
<entities>
<LOCATIONS>
{{LOCATIONS_XML}}
</LOCATIONS>

<PERSONS>
{{PERSONS_XML}}
</PERSONS>

<ORGANIZATIONS>
{{ORGANIZATIONS_XML}}
</ORGANIZATIONS>

<PRODUCTS>
{{PRODUCTS_XML}}
</PRODUCTS>

<ARTS>
{{ARTS_XML}}
</ARTS>

<EVENTS>
{{EVENTS_XML}}
</EVENTS>

<BUILDINGS>
{{BUILDINGS_XML}}
</BUILDINGS>

<MISCELLANEOUS>
{{MISCELLANEOUSS_XML}}
</MISCELLANEOUS>
</entities>

2. A history of fictional events involving these entities:
<history>
{{HISTORY_XML}}
</history>

3. The date of the next fictional event: {{DATE}}

3. An outline describing the next fictional event:
<outline>
{{OUTLINE}}
</outline>

Follow these steps to complete your task:

1. Carefully compare the events described in the outline with the events from the history. Look for any inconsistencies or contradictions.

2. Compare the named entities described in the outline with the list of provided named entities. Ensure they are consistent.

3. Note that changes to known named entities are acceptable if they are reasonably discussed within the outline.

4. If you find any inconsistencies or contradictions:
a. Make minimal changes to the outline to resolve the issues.
b. Ensure your changes maintain the original structure and flow of the outline as much as possible.
c. Double-check after fixing the inconsistencies to ensure they no longer persist.

5. If the outline is consistent with the provided list of entities and previous events (i.e., not contradictory), output the outline without any changes.

Present your results in the following format:

```
<results>
<scratchpad>
(Your reasoning process, including any inconsistencies found and how you resolved them)
</scratchpad>
<storyitems>
<storyitem>[First story item from the outline, corrected if necessary]</storyitem>
<storyitem>[Second story item from the outline, corrected if necessary]</storyitem>
<storyitem>[Continue with additional story items as needed]</storyitem>
</storyitems>
</results>
```

Important reminders:
- Always refer to named entities using their full name as described by the "name" property.
- Remember that all events are fictional but should sound plausible and realistic.
- Your output must include all storyitems, whether you had adjusted them or not.
- Do not add new events or significantly alter the plot. Your task is to ensure consistency, not to rewrite the story.

If you cannot resolve an inconsistency without significantly changing the plot, note this in your scratchpad and leave the story item as is, highlighting the inconsistency.

Begin your analysis now, and provide your results in the format specified above.

H.5 Event Outline (novel) Named Entity Recognition Prompt

You are an AI assistant tasked with identifying new named entities from a given outline of a fictional event description. Your goal is to identify any new named entities mentioned in the OUTLINE that are not already present in the provided lists of existing entities.

First, carefully read and analyze the following OUTLINE:

```
Date: {{DATE}}
<OUTLINE>
{{OUTLINE}}
</OUTLINE>
```

Now, review the existing named entities in the following XML structures. Note that these lists may be empty or contain partial information:

```
<entities>
{{LOCATIONS_XML}}
{{PERSONS_XML}}
{{ORGANIZATIONS_XML}}
{{PRODUCTS_XML}}
{{ARTS_XML}}
{{EVENTS_XML}}
{{BUILDINGS_XML}}
{{MISCELLANEOUSS_XML}}
</entities>
```

The definitions with examples for each named entity category are listed here:
- Person
Definition: Individual human beings, including fictional figures.
Examples: Barack Obama, William Shakespeare, Harry Potter, Marie Curie

-Organization
Definition: Groups of people working together for a common purpose, including companies, institutions, ethnic groups, communities and government bodies.
Examples: United Nations, Apple Inc., Harvard University, Greenpeace, Maori

- Location
Definition: Geographical or political areas, locations, countries (together with their nationalities), cities, natural landmarks, and regions.
Examples: Paris, Mount Everest, California, Amazon Rainforest, Japan, Australia, Australian

- Product
Definition: Goods or services created for consumer use or commercial purposes.
Examples: iPhone, Coca-Cola, Microsoft Office, Tesla Model 3

- Art
Definition: Creative works in various forms, including artifacts, ornaments, visual arts, literature, music, and performance.
Examples: Mona Lisa, To Kill a Mockingbird, Beethoven's Symphony No. 9, Hamilton (musical)

- Building
Definition: Structures designed for human occupancy or use, including residential, commercial, and public structures.
Examples: Empire State Building, Taj Mahal, Sydney Opera House, Buckingham Palace

- Event
Definition: Significant occurrences or planned gatherings, including historical moments, celebrations, and competitions.
Examples: World War II, Olympic Games, Woodstock Music Festival, Super Bowl

- Miscellaneous
Definition: Other named entities that don't fit into the above categories, such as abstract concepts, unique identifiers.
Examples: Theory of Relativity, Morse Code, Brexit, Zodiac Signs

IMPORTANT: You must ONLY identify a named entity if the outline explicitly refers to the entity by name.
DO NOT list entities for events which are not explicitly referred to by name.
DO NOT list entities for buildings that are not explicitly referred to by name (e.g. a big Chicago villa)

Example: "On the 27th birthday of Carla Short"
- "Carla Short" is a named entity (Person)
- No named event is in this statement

Example: "On the third day of the Banana Split Festival I left my keys"
- National Banana Split Festival is a named entity (Event)

Example: "I sold my 13$MChicagovilla.”$
$-”Chicago”isanamedentity(Location)$
$-$”13M Chicago villa" is NOT a named entity. However, this phrase contains the named entity Chicago (location)

Example: "I bought tickets to go on the Sky Needle, the highest skyscraper in town!"
- "Sky Needle" is a named entity (building)

Example: "Aboriginal people from a well known Australian city."
- "Aboriginal" is a named entity (Organization)
- "Australia" is a named entity (Location)

To complete this task, follow these steps:

1. Identification of named entities: Carefully read through the OUTLINE and identify all named entities (locations, persons, organizations, products, art, events, buildings, or miscellaneous) that are mentioned by name. Consider all named entities. Carefully follow the definition and examples provided for each entity type above. DO NOT ignore named entities from the outline that exist in the real world. Carefully double-check that you did not miss any named entities. If you are in doubt about some named entities, epxlain why you bare not sure.

2. Remove entities that are already known and only keep new named entities:
a. For each named entity you identify, check if it already exists in the <entities>. If <entities> is empty, consider all

named entities as new.
b. Only keep new named entities which cannot be found in the <entities> list of known named entities.
c. Ensure that each new named entity is explicitly mentioned by name in the <outline>. The outline may refer to various entities (such as events, buildings, products) that are described generally but not by a specific name. These should not be identified as named entities.

3. Verification and refinement of new named entities:
a. Review your preliminary list of new named entities.
b. Double-check each named entity against the existing XML structures to ensure it is truly new and not already present.
c. For each new named entity, determine its full name as it appears in the OUTLINE. Do not infer or create additional information.
d. Ensure that the full name of each named entity is distinct. Two different entities MUST NOT share the same full name.
e. Examine your list of verified new named entities for any redundancies. If any entities refer to the same person, place, organization, product, art piece, event, building, or miscellaneous item but are mentioned with slight variations in the OUTLINE, keep only the most complete or accurate version and remove the others.
f. Do not list abbreviations as separate entities. If an abbreviation is used, do not include it with the full name of the entity, and do not create a separate entry for it.
g. Ensure that each identified named entity refers to an actual identifiable entity mentioned in the OUTLINE.

4. Formatting and outputting the final list:
a. Categorize each verified new named entity as either a location, person, organization, product, art, event, building, or miscellaneous.
b. Format your output for each new named entity using the appropriate XML tag based on its category:
- For locations: <location>Full Name of Location</location>
- For persons: <person>Full Name of Person</person>
- For organizations: <organization>Full Name of Organization</organization>
- For products: <product>Full Name of Product</product>
- For art: <art>Full Name of Art Piece</art>
- For events: <event>Full Name of Event</event>
- For buildings: <building>Full Name of Building</building>
- For miscellaneous: <miscellaneous>Full Name of Miscellaneous Entity</miscellaneous>

c. If no new named entities are found after verification and redundancy removal, output <no_new_entities>No new named entities identified</no_new_entities>

5. Double-check miscellaneous
a. If you classified any new entity as miscellaneous:
- Compare the entity and how it is used in context with all other named entity types
- Check if any of the other entity types fits this named entity (it does not need to fit 100- If any of these entities seem applicable to the entity, list the named entity with the newly found entity and NOT as miscellaneous entity.

Sometimes the difference between the named entities is blurry. To help you make decisions in borderline cases follow these rules:
- Organization or Building: If in doubt, use "organization" instead of "building". Only use "building" if the name alone clearly identifies the entity as a building.

Important notes:
- All entities in this task should be fictional and should be treated as distinct from any existing real-world or fictional entities. However, you must identify the names exactly as they appear in the OUTLINE, even if they are similar to real-world entities.
- Remember not to list abbreviations as separate entities.
- Include the names of any real-world entities from the outline.
- Only list NAMED entities. Do not create names for all entities that occur within the story.
- Make sure that you have an output for EACH NEW named entity that you have identified!

Use a <scratchpad> node within the <results> root node for all your reasoning. DO NOT include any XML tags in the reasoning process.

Present your final list of new named entities in this format:

<output>
(Your list of new named entities in XML format goes here)
</output>

If you need to think through your process or make notes, use UPPERCASE variable names for your thinking process. Your final output should only include the new named entities or the no_new_entities tag if applicable.

List all output including your thinking process within a single <results> node.

Before finalizing your output, double-check that you did not miss any named entities from the outline and that each named entity has a distinct full name.

H.6 Named Entity Name Generation Prompt

You are an AI assistant tasked with renaming a list of entities to ensure they are distinct from any known real-world or fictional names. This task is crucial for creating original content that doesn't infringe on existing intellectual property or cause confusion with real entities.

You will be provided with eight lists of names to be renamed: locations, organizations, persons, products, art, buildings, events, and miscellaneous items. These are presented in XML format as follows:

Entity names to change:

```
<entity_names>
{{LOCATIONS_NAME_XML}}
{{ORGANIZATIONS_NAME_XML}}
{{PERSONS_NAME_XML}}
{{PRODUCTS_NAME_XML}}
{{ARTS_NAME_XML}}
{{BUILDINGS_NAME_XML}}
{{EVENTS_NAME_XML}}
{{MISCELLANEOUSS_NAME_XML}}
</entity_names>
```

Follow these steps to complete the task:

1. For each entity in the lists, create a new name that is different from the original but maintains a similar style or feel. The new name must be fictional, but it must sound realistic. Avoid names that are clearly fictional.

2. Ensure that the new names are not associated with any known real-world or fictional entities. This includes names of people, places, organizations, products, artworks, buildings, events, or characters from books, movies, or other media.

3. When creating new names:
- For locations: Maintain a geographical feel appropriate to the original name's region.
- For organizations: Keep a professional or institutional tone similar to the original.
- For persons: Preserve the cultural or ethnic flavor of the original name if applicable. If only a first name is provided, consider adding the lastname. If only a first name is provided, consider adding a last name. Ensure consistency in naming, particularly with inter-personal relations. For example, children should have the same last name as their parents, and married people often share the same last name.
- For products: Retain a similar market appeal and product category feel.
- For art: Maintain the artistic style or genre suggested by the original name.
- For buildings: Keep architectural or functional implications of the original name.
- For events: Preserve the nature or purpose of the event in the new name.
- For miscellaneous items: Retain the essence or category of the original item.

4. Avoid using common words, phrases, or combinations that might accidentally reference existing entities.

5. For each renamed entity, provide both the new name and the old name.

6. Output your results in the following XML format:

```
<results>
<location>
<name>[New Location Name]</name>
<old_name>[Original Location Name]</old_name>
</location>
<organization>
<name>[New Organization Name]</name>
<old_name>[Original Organization Name]</old_name>
</organization>
<person>
<name>[New Person Name]</name>
<old_name>[Original Person Name]</old_name>
</person>
```

```
<product>
<name>[New Product Name]</name>
<old_name>[Original Product Name]</old_name>
</product>
<art>
<name>[New Art Name]</name>
<old_name>[Original Art Name]</old_name>
</art>
<building>
<name>[New Building Name]</name>
<old_name>[Original Building Name]</old_name>
</building>
<event>
<name>[New Event Name]</name>
<old_name>[Original Event Name]</old_name>
</event>
<miscellaneous>
<name>[New Miscellaneous Item Name]</name>
<old_name>[Original Miscellaneous Item Name]</old_name>
</miscellaneous>
</results>
```

Here’s an example of how your output should look:

```
<results>
<location>
<name>Emerald Heights</name>
<old_name>Green Valley</old_name>
</location>
<organization>
<name>Quantum Dynamics Corporation</name>
<old_name>Particle Physics Institute</old_name>
</organization>
<person>
<name>Zara Blackwood</name>
<old_name>Emma Stone</old_name>
</person>
<product>
<name>SonicWave X1</name>
<old_name>iPod</old_name>
</product>
<art>
<name>Whispers of Eternity</name>
<old_name>Starry Night</old_name>
</art>
<building>
<name>Pinnacle Tower</name>
<old_name>Empire State Building</old_name>
</building>
<event>
<name>Global Harmony Festival</name>
<old_name>Woodstock</old_name>
</event>
<miscellaneous>
<name>LumiGlow Orb</name>
<old_name>Lava Lamp</old_name>
</miscellaneous>
</results>
```

Remember to create unique names for each entity and ensure they don’t match any known entities. Provide your complete list of renamed entities in the specified XML format. When coming up with new names, try to keep them sounding realistic and appropriate to their category (location, organization, person, product, art, building, event, or miscellaneous).

Begin your renaming process now, and present your results in the format specified above. You may use a <scratchpad> section within the <results> tags to show your thinking process if needed. All actual results should be listed in <location>, <organization>, <person>, <product>, <art>, <building>, <event>, or <miscellaneous> nodes within the <results> node.

Do not include any content outside of the <results> tags in your response.

ONLY change the names of the entities listed under entity names to change. Do not invent any other entities. Do not change the names of any other entities.

H.7 Event Outline Adjustment with new Named Entity Names

You are an AI assistant tasked with updating an OUTLINE to be consistent with new entity names. Your goal is to make minimal changes while ensuring all entity names are updated correctly. Follow these instructions carefully:

First, here are the entities for which the names have been changed:

```
<entities>
<adjusted_locations>
{{ADJUSTED_LOCATIONS_XML}}
</adjusted_locations>
<adjusted_persons>
{{ADJUSTED_PERSONS_XML}}
</adjusted_persons>
<adjusted_organizations>
{{ADJUSTED_ORGANIZATIONS_XML}}
</adjusted_organizations>
<adjusted_products>
{{ADJUSTED_PRODUCTS_XML}}
</adjusted_products>
<adjusted_arts>
{{ADJUSTED_ARTS_XML}}
</adjusted_arts>
<adjusted_buildings>
{{ADJUSTED_BUILDINGS_XML}}
</adjusted_buildings>
<adjusted_events>
{{ADJUSTED_EVENTS_XML}}
</adjusted_events>
<adjusted_miscellaneouss>
{{ADJUSTED_MISCELLANEOUSS_XML}}
</adjusted_miscellaneouss>
</entities>
```

Your task is to update the OUTLINE to be consistent with the new names of these entities. Follow these instructions carefully:

1. Make minimal changes to the outline. Only update the names of entities that have been changed.
2. Apply changes on each sentence individually.
3. Output each updated sentence as a separate <storyitem>.
4. If a sentence does not contain any entities that need to be changed, output it as is.
5. Ensure that you maintain the original structure and content of the OUTLINE, changing only the necessary entity names.
6. Always use the full name as defined by the "name" property of the entities.

Output format:
Place all your outputs in a root node <results>. Do not output any content outside of this root node. Each sentence should be in its own <storyitem> tag. Your output should look like this:

```
<results>
<storyitem>[Updated sentence 1]</storyitem>
<storyitem>[Updated sentence 2]</storyitem>
...
<storyitem>[Updated sentence n]</storyitem>
</results>
```

Important reminders:
- Make only the necessary changes to reflect the new entity names while preserving the original meaning and structure of each sentence.
- Ensure that all entity name changes are consistent with the provided data for locations, persons, organizations, products, art, buildings, events, and miscellaneous items.

- Double-check your work to make sure you haven't missed any entity name changes or accidentally modified any content that should remain unchanged.
- Change a name only if it refers to the specific named entity being updated. Do not change the name if it refers to a different entity, even if their names partially overlap.

Now, here is the OUTLINE to update:

Date: {DATE}
<outline>
{OUTLINE}
</outline>

Process each sentence in the OUTLINE, updating entity names as necessary, and output the results as instructed above.

DO NOT include the date as a storyitem.
Ensure that you DO NOT change the names of any previously identified entities whose names did not require modification.

H.8 Event Outline (all) Named Entity Recognition Prompt

You are an AI assistant tasked with identifying which of the provided entities are explicitly named within the provided outline. Follow these instructions carefully:

1. The date for this fictional event outline is:
<date>{{DATE}}</date>

2. Here is the outline you need to analyze:
<outline>
{{OUTLINE}}
</outline>

3. Here is the list of known (fictional) named entities:
<entities>
{{LOCATIONS_XML}}
{{PERSONS_XML}}
{{ORGANIZATIONS_XML}}
{{PRODUCTS_XML}}
{{ARTS_XML}}
{{EVENTS_XML}}
{{BUILDINGS_XML}}
{{MISCELLANEOUSS_XML}}

{{ADJUSTED_LOCATIONS_XML}}
{{ADJUSTED_PERSONS_XML}}
{{ADJUSTED_ORGANIZATIONS_XML}}
{{ADJUSTED_PRODUCTS_XML}}
{{ADJUSTED_ARTS_XML}}
{{ADJUSTED_BUILDINGS_XML}}
{{ADJUSTED_EVENTS_XML}}
{{ADJUSTED_MISCELLANEOUSS_XML}}
</entities>

4. Your task is to go through all of the named entities in the provided list. For each named entity, check if it is explicitly referred to by name in the outline. If a named entity is explicitly mentioned by name, include it in your results.

5. For each entity you identify, list them using the following format:
<[entity_type]><id>[id of the entity]</id><name>[full name of the entity]</name></[entity_type]>

6. Present your final output within a single <results> root node, structured as follows:

<results>
<entities>
(List all identified entities here as described in step 5)
</entities>
</results>

7. Double-check your work: For each entity that you have identified, make sure to find the sentence within the outline that explicitly refers to this identified entity by name. If you cannot find that the entity is mentioned explicitly by name in the outline, do not include it in your results.

8. You can use a <scratchpad> for your thinking process. Do not include any XML tags within your scratchpad.

9. Make sure that you have not missed any named entity from the outline that was also provided to you in the list of named entities.

Remember, only include entities that are explicitly mentioned by name in the outline. Do not infer or assume the presence of entities that are not directly named.

H.9 Event Outline Named Entity Resolution Prompt

You are an AI assistant tasked with identifying and marking fictional entities within an outline of a fictional event. Your goal is to replace each occurrence of these entities with a specific format that includes the entity's ID. Follow these instructions carefully:

You will be working with the following inputs:

Date: {{DATE}}

<outline>
{{OUTLINE}}
</outline>

<entities>
{{ALL_USED_LOCATION_XML}}
{{ALL_USED_PERSON_XML}}
{{ALL_USED_ORGANIZATION_XML}}
{{ALL_USED_PRODUCT_XML}}
{{ALL_USED_ART_XML}}
{{ALL_USED_BUILDING_XML}}
{{ALL_USED_EVENT_XML}}
{{ALL_USED_MISCELLANEOUS_XML}}
</entities>

Process the outline by following these steps:

1. Carefully review the list of entities provided in the <entities> section. Each entity will have an associated ID and a name.

2. Search the outline for all occurrences of each entity in the list.

3. For each occurrence found, replace it with the format: {full name|ID}
Where "full name" is the full name as provided via the "name" property of the entity.

4. Maintain the original structure and formatting of the outline, only changing the entities as described.

5. After processing all entities, review the entire outline to ensure all occurrences have been properly marked and no entities were missed.

6. Output the processed outline, maintaining its original structure but with all entity occurrences replaced as instructed.

Important points to remember:
- Be thorough in your search for entities, including variations or partial mentions.
- Always use the ID and name as provided in the entity list, even:
- If an entity is referred to by full name
- If an entity is referred to by an abbreviation
- If an entity is referred to using parts of the full name

Format your output as follows:
- Enclose the entire processed outline within <outline> tags.
- Place each sentence of the outline within separate <storyitem> tags.

Provide your final output without any additional commentary or explanations. Focus solely on processing the outline as instructed.

H.10 Named Entity KB Entry Generation Prompt

You are an AI assistant tasked with creating fictional entities based on provided information.
Your goal is to generate detailed, coherent, and realistic descriptions for new locations, persons, organizations, products, art, buildings, events, and miscellaneous entities.
Follow these instructions carefully:

1. Review the existing entities (if provided):

<existing_entities>
{{LOCATIONS_XML}}
{{PERSONS_XML}}
{{ORGANIZATIONS_XML}}
{{PRODUCTS_XML}}
{{ARTS_XML}}
{{EVENTS_XML}}
{{BUILDINGS_XML}}
{{MISCELLANEOUSS_XML}}
</existing_entities>

2. Review the names of new entities to be created:
<new_entity_names>
{{USED_NEW-LOCATIONS_XML}}
{{USED_NEW-PERSONS_XML}}
{{USED_NEW-ORGANIZATIONS_XML}}
{{USED_NEW-PRODUCTS_XML}}
{{USED_NEW-ARTS_XML}}
{{USED_NEW-EVENTS_XML}}
{{USED_NEW-BUILDINGS_XML}}
{{USED_NEW-MISCELLANEOUSS_XML}}
</new_entity_names>

3. Carefully read the provided outline:

Date: {{DATE}}
<outline>
{{OUTLINE}}
</outline>

Based on the information provided, create detailed descriptions for each new entity following these guidelines:

General instructions:
- Ensure all created entities are entirely fictional and not similar to any real or known fictional entities.
- Maintain realism and coherence with the provided outline and other entities.
- Create a believable and consistent fictional world that aligns with the context of the outline.
- For entity descriptions, focus on providing a solid background that remains valid throughout the story, rather than basing it centrally on the outline itself.
- Develop well-rounded entities with backgrounds and characteristics that can support various potential story developments beyond the specific outline provided.
- Strictly derive all entities from the outline. Do not invent entities that are not mentioned in the outline.
- Do not alter the name of the entities.
- Some properties (e.g., place, city, country, spouse, architect, country, nationality) must be filled with named entities. Make sure to use FICTIONAL named entities (fictional places, cities, countries, spouse names, architects, countries, nationalities, etc.). Check if you should use one of the existing fictional named entities (from <new_entity_names> and <existing_entities>) or create a fictional name instead. DO NOT say that any of the entities or properties are fictional. It is important that everything seems realistic!
- Avoid exaggerating the named entities. They can be ordinary and don't need to be world-class or state-of-the-art.
- Ensure the details of the named entities are realistic. If global impact isn't necessary, adjust the details to be more modest and appropriate.

Specific instructions for each entity type:

1. For each new location:

```
- Use the provided name and ID
- Determine an appropriate type (city, village, country, region, etc.)
- Write a concise single-sentence description that:
* Does not refer to any other entities
* Provides only background information, not related to the event in the outline
* Is general and serves as background for this event
- Include at least five additional properties from: population, area, founded, climate, elevation, country
- Format the output as follows:
<location>
<id>[Provided ID]</id>
<name>[Fictional name]</name>
<type>[city/village/country/region]</type>
<description>[One-sentence concise description]</description>
(Include at least 5 of: population, area, founded, climate, elevation, country)
</location>

2. For each new person:
- Use the provided name and ID
- Create fictional details for: date_of_birth, gender, profession, nationality, education
- Write a concise single-sentence description that:
* Focuses on background, personality, and motivations
* Does not refer to any other entities
* Provides only background information, not related to the event in the outline
* Is general and serves as background for this event
- Include at least five additional properties from: height, weight, eye_color, hair_color, political_affiliation, marital_status, spouse
- Format the output as follows:
<person>
<id>[Provided ID]</id>
<name>[Full name]</name>
<date_of_birth>[Date]</date_of_birth>
<gender>[Gender]</gender>
<profession>[Job title]</profession>
<nationality>[Country]</nationality>
<education>[Highest level of education]</education>
<description>[One-sentence concise description]</description>
(Include at least 5 of: height, weight, eye_color, hair_color, political_affiliation, marital_status)
</person>

3. For each new organization:
- Use the provided name and ID
- Determine an appropriate type (company, non-profit, educational institution, government agency, etc.)
- Write a concise single-sentence description that:
* Does not refer to any other entities
* Provides only background information, not related to the event in the outline
* Is general and serves as background for this event
- Include at least five additional properties from: founded, headquarters, industry, mission_statement, number_of_employees, annual_revenue
- Format the output as follows:
<organization>
<id>[Provided ID]</id>
<name>[Fictional name]</name>
<type>[company/non-profit/educational institution/government agency]</type>
<description>[One-sentence concise description]</description>
(Include at least 5 of: founded, headquarters, industry, mission_statement, number_of_employees, annual_revenue)
</organization>

4. For each new product:
- Use the provided name and ID
- Determine an appropriate type (consumer good, software, service, etc.)
- Write a concise single-sentence description that:
* Does not refer to any other entities
* Provides only background information, not related to the event in the outline
* Is general and serves as background for this event
- Include at least five additional properties from: manufacturer, release_date, price, weight, warranty
- Format the output as follows:
<product>
<id>[Provided ID]</id>
```

```
<name>[Fictional name]</name>
<type>[consumer good/software/service]</type>
<description>[One-sentence concise description]</description>
(Include at least 5 of: manufacturer, release_date, price, weight, warranty)
</product>

5. For each new art piece:
- Use the provided name and ID
- Determine an appropriate type (painting, sculpture, novel, film, etc.)
- Write a concise single-sentence description that:
* Does not refer to any other entities
* Provides only background information, not related to the event in the outline
* Is general and serves as background for this event
- Include at least five additional properties from: creator, year_created, current_location_country, current_location_city, current_location_place
- Format the output as follows:
<art>
<id>[Provided ID]</id>
<name>[Fictional name]</name>
<type>[painting/sculpture/novel/film]</type>
<description>[One-sentence concise description]</description>
(Include at least 5 of: creator, year_created, current_location_country, current_location_city, current_location_place)
</art>

6. For each new building:
- Use the provided name and ID
- Determine an appropriate type (residential, commercial, public, etc.)
- Write a concise single-sentence description that:
* Does not refer to any other entities
* Provides only background information, not related to the event in the outline
* Is general and serves as background for this event
- Include at least five additional properties from: place, city, country, architect, year_built, height, floors, material, capacity
- Format the output as follows:
<building>
<id>[Provided ID]</id>
<name>[Fictional name]</name>
<type>[residential/commercial/public]</type>
<description>[One-sentence concise description]</description>
(Include at least 5 of: place, city, country, architect, year_built, height, floors, material, capacity)
</building>

7. For each new event:
- Use the provided name and ID
- Determine an appropriate type (historical, cultural, sporting, etc.)
- Write a concise single-sentence description that:
* Does not refer to any other entities
* Provides only background information, not related to the event in the outline
* Is general and serves as background for this event
- Include at least five additional properties from: date, place, city, country, duration, organizer, number_of_participants, budget
- Format the output as follows:
<event>
<id>[Provided ID]</id>
<name>[Fictional name]</name>
<type>[historical/cultural/sporting]</type>
<description>[One-sentence concise description]</description>
(Include at least 5 of: date, place, city, country, duration, organizer, number_of_participants, budget)
</event>

8. For each new miscellaneous entity:
- Use the provided name and ID
- Determine an appropriate type (concept, theory, phenomenon, etc.)
- Write a concise single-sentence description that:
* Does not refer to any other entities
* Provides only background information, not related to the event in the outline
* Is general and serves as background for this event
- Format the output as follows:
```

```
<miscellaneous>
<id>[Provided ID]</id>
<name>[Fictional name]</name>
<type>[concept/theory/phenomenon]</type>
<description>[One-sentence concise description]</description>
</miscellaneous>

Present your final output in the following format:
<results>
(Insert all created entities here, grouped by type)
</results>

Remember to create detailed, coherent, and realistic descriptions for each entity while adhering to the guidelines provided. Ensure that all descriptions are single, short, and concise sentences that do not refer to other entities and only discuss background information unrelated to the event described in the outline.
```

H.11 Named Entity Update Prompt

```
You are an AI assistant tasked with updating a list of fictional entities based on an outline of events. Follow these instructions carefully to complete the task:

1. First, read the outline of the fictional event:
<outline>
{{OUTLINE}}
</outline>

2. Next, review the list of fictional entities:
<entities>
{{USED_LOCATION_XML}}
{{USED_PERSON_XML}}
{{USED_ORGANIZATION_XML}}
{{USED_PRODUCT_XML}}
{{USED_ART_XML}}
{{USED_EVENT_XML}}
{{USED_BUILDING_XML}}
{{USED_MISCELLANEOUS_XML}}
</entities>

3. For each entity in the list, follow these steps:
a. Identify the entity's role in the outline. Create an update sentence describing how the entity was affected by or involved in the events described in the outline.
b. Review all properties of the entity EXCEPT for "id", "created_at", "history", and "entity_class".
c. For each property:
- If the property does not need to be updated based on the outline, leave it as-is.
- If the events or the time difference since the last update (last_updated) indicate that the value has changed, update it accordingly.
- Ensure the new value is consistent with the outline, other entities, and plausible given the time difference.
d. Do not alter the "description" field unless the current description is no longer valid after the events in the outline.

4. Output your results for each entity in the following format:
<results>
<[entity_type]>
<entity_id>[Insert entity id here]</entity_id>
<update>[Insert your update sentence here]</update>
(List any properties that were changed, with their new values, using the format:)
<[property_name]>[new value]</[property_name]>
</[entity_type]>
(Repeat for each entity)
</results>

5. Important reminders:
- Stick to the information provided in the outline and entities list. Do not invent new details or events.
- Ensure all updates and changes are consistent with the outline and with each other.
- Be concise in your updates, focusing only on relevant changes.
- If no properties need to be changed for an entity, do not include any property tags.
- Process each entity in the order they are presented in the list.
- Use the appropriate entity type tag (e.g., <location>, <person>, etc.) instead of <entity_update>.
```

- Begin your response with the first entity update immediately, without any preamble.

All actual results should be listed in <location>, <organization>, <person>, <product>, <art>, <building>, <event>, or <miscellaneous> nodes within the <results> node.
Do not include any content outside of the <results> tags in your response.

Start processing the entities now, following the instructions and format provided above.

H.12 Diverse Next Summary Generation Prompt

You are an AI tasked with creating fictional future news summaries based on provided information. Your goal is to generate plausible continuations of an existing narrative. Follow these instructions carefully:

1. First, you will be given known entities in this fictional world. These will be provided in the following format:
<known_entities>
{{LOCATIONS_XML}}
{{PERSONS_XML}}
{{ORGANIZATIONS_XML}}
{{PRODUCTS_XML}}
{{ARTS_XML}}
{{EVENTS_XML}}
{{BUILDINGS_XML}}
{{MISCELLANEOUSS_XML}}
</known_entities>

2. Next, you will be provided with the history of events that have already occurred in this fictional world:

<history>
{{HISTORY_XML}}
</history>

3. Your task is to create {{NUMBER_SUMMARIES}} new summaries that describe future fictional events following the last event from the <history>. These new summaries should be consistent with the existing story and represent plausible continuations or developments of the original narrative.

4. For each summary, create:
a. A summary text (a single concise sentence)
b. The date on which this fictional next event happens

5. Before starting, check if the history of events indicates specific dates for followup events. Ensure your continuations are consistent with these expected followup events. All your summaries must either concern this event, or happen before this event.

6. Make sure that each summary you generate focuses on at least one of the main named entities from the history of events.

7. Follow these guidelines when creating your summaries:
a. Ensure all summaries are fictional and not based on real events or real people.
b. Make the summaries sound realistic and plausible as follow-up stories to the previous outlines.
c. Think about plausible next events based on the fictional named entities, the history of the past fictional events and the genre {{GENRE}}.
d. Create summaries that are unbiased and objective in tone.
e. Each summary MUST BE ONLY A SINGLE concise sentence.
f. Summaries may focus on different personas or organizations from the provided lists.
g. Ensure a balance of positive and negative news stories, developments, and alternative scenarios.
h. Consider various dimensions or personas that could be varied when generating diverse summaries.

Try to be diverse in the summaries you generate. Consider different plausible substories and vary:
- between impact: Try to create various low-impact next events, but you may also at times mix in an event with a slightly higher impact.
- between directions: Vary between positive and negative story developments. Think about how stories in the {{GENRE}} genre progress in the real-world, NOT in a novel. Provide various realistic alternatives for how the story may progress in either direction.
- between different key named entities of the fictional story.

8. Output format:

- Enclose each summary in <summary> tags.
- Each summary must have two child properties:
<text>[The generated summary]</text>
<date>[The date for the next event]</date>
- Before each summary, explain your thought process in <thought_process> tags. Make sure to identify all known followup events based on the provided history first, and verify that your continuations are consistent with these known followup events regarding the date.
- Output everything within a <results> root node.

9. Special instructions:
- The history and entities have special formatting. They sometimes include statements like {[PHRASE]|[ID]}. Read it as [PHRASE], while the [ID] specifies the ID of the linked entity.
- Example: "I met {Boris Bowman|PERSON-1} yesterday." Read as: "I met Boris Bowman yesterday." (The ID of Boris Bowman is PERSON-1).
- DO NOT use the {[PHRASE]|[ID]} formulation when generating new summaries.
- Do not exaggerate the summaries. Avoid using words like "groundbreaking", "worldwide", "global". Keep the summaries realistic.
- Do not create summaries with global or national impact unless the genre specifically requires it. Instead, focus on smaller or local developments.
- Do not focus on technological discoveries or topics like AI tools, virtual reality, augmented reality, 3D-modelling, quantum computing, etc. You may include such topics only if they are HIGHLY relevant to the genre {{GENRE}} AND the provided history of events.
- Focus on realistic, meaningful summaries with specific details and developments that align with typical, realistic scenarios of the genre {{GENRE}}.

10. Final reminders:
- Repeat this process for all {{NUMBER_SUMMARIES}} summaries.
- Ensure that each summary explores a different aspect or potential next step of the fictional situation presented in the HISTORY.
- Each summary text MUST BE ONLY A SINGLE concise sentence.
- The continuations should be diverse in terms of high (at most one), medium and low impact, positive and negative story developments, etc., and can cover different alternatives of how the future event can play out.
- Do not enumerate over the summaries.
- When brainstorming future event summaries, carefully consider whether each continuation aligns with the provided history of events, provided named entities, and the genre {{GENRE}}. Before creating summaries with global or large-scale impact, double-check if such developments seem plausible based on how real-world events of this type would typically unfold. Focus on what fits the genre and provided history, ensuring that every dimension feels realistic and consistent with the context.

Remember to maintain the desired format and brevity of the event summaries while creating plausible and engaging continuations of the narrative.

Here's an example of the desired output format:

<results>
<thought_process>
(This includes the thought process for all summaries.)
</thought_process>

<summaries>
<summary>
<text>[Your first summary text here]</text>
<date>[Date for the first summary]</date>
</summary>

<summary>
<text>[Your second summary text here]</text>
<date>[Date for the second summary]</date>
</summary>

(... continue for all summaries ...)
</summaries>
</results>

H.13 Mutually Exclusive Summary Generation Prompt

You are an AI tasked with creating alternative fictional future news summaries based on provided information.
Your goal is to generate summaries of plausible alternative continuations of an existing narrative. Follow these instructions carefully:

1. First, you will be given known entities in this fictional world:

<known_entities>
{{LOCATIONS_XML}}
{{PERSONS_XML}}
{{ORGANIZATIONS_XML}}
{{PRODUCTS_XML}}
{{ARTS_XML}}
{{EVENTS_XML}}
{{BUILDINGS_XML}}
{{MISCELLANEOUSS_XML}}
</known_entities>

2. Next, you will be provided with the history of events that have already occurred in this fictional world:

<history>
{{HISTORY_XML}}
</history>

3. You will be given the following plausible summary of how the fictional event evolves:
<continuation_summary>
<text>{{CONTINUATION_TEXT}}</text>
<date>{{CONTINUATION_DATE}}</date>
</continuation_summary>

4. Your task is to create {{NUM_ALTERNATIVES}} contradictory alternative summaries of how the fictional event can progress based on the provided continuation summary.
Each of these continuation summaries must make subtle changes to the continuation summary such that they are contradictory alternatives to one another.
This means, if the story evolves with one of the continuation summaries, the other ones cannot happen anymore.

5. When changing the provided continuation summary, maintain these key properties:
a) The central topic and involved main entity
b) The stance (whether this is a positive, neutral or negative story evolvement)
c) The impact (whether this is a high impact, medium impact or low impact evolvement)
d) The same date

6. Guidelines for creating alternative summaries:
- Ensure each new continuation summary is consistent with the existing story and represents a plausible continuation or development of the original narrative.
- Write each continuation summary as a single concise sentence in an objective tone.
- Make sure your continuations are consistent with any known followup events regarding the date.

7. Output format:
- Enclose your entire response in <results> tags.
- Before the summaries, explain your thought process in <thought_process> tags. Make sure to identify all known followup events based on the provided history first, and verify that your continuations are consistent with these known followup events regarding the date.
- Enclose all summaries in <summaries> tags.
- For each summary:
<summary>
<text>[The generated summary]</text>
<date>[The date for the next event]</date>
</summary>

8. Special instructions:
- The history and entities have special formatting. They sometimes include statements like {[PHRASE]|[ID]}. Read it as [PHRASE], while the [ID] specifies the ID of the linked entity.
- Example: "I met {Boris Bowman|PERSON-1} yesterday." Read as: "I met Boris Bowman yesterday." (The ID of Boris Bowman is PERSON-1).
- DO NOT use the {[PHRASE]|[ID]} formulation when generating new summaries.
- When brainstorming alternative event summaries, carefully consider whether each continuation aligns with the provided history of events, provided named entities, and the genre {{GENRE}}. Before creating summaries with global or large-scale impact, double-check if such developments seem plausible based on how real-world events of this

type would typically unfold. Focus on what fits the genre and provided history, ensuring that every dimension feels realistic and consistent with the context.
- Do not exaggerate the summaries. Avoid making summaries "groundbreaking." Keep the summaries realistic.
- Do not create summaries with global or national impact unless the genre specifically requires it. Instead, focus on smaller or local developments.
- Do not focus on technological discoveries or topics like AI tools, virtual reality, quantum computing, etc. You may include such topics only if they are HIGHLY relevant to the genre {{GENRE}} AND the provided history of events.
- Focus on realistic, meaningful summaries with specific details and developments that align with typical, realistic scenarios of the genre {{GENRE}}.

9. Here's an example of the desired output format:

<results>
<thought_process>
(Your explanation of the thought process, including identification of known followup events and verification of consistency)
</thought_process>
<summaries>
<summary>
<text>[First alternative summary]</text>
<date>[Date of the event]</date>
</summary>
<summary>
<text>[Second alternative summary]</text>
<date>[Date of the event]</date>
</summary>
(Additional summaries as needed)
</summaries>
</results>

Remember to maintain the desired format and brevity of the event continuations while creating plausible and engaging continuations of the narrative. Begin your response with the <results> tag and end it with the </results> tag.

H.14 Sentence Selection Prompt for Time-span Questions

You are an AI assistant tasked with analyzing fictional event descriptions and identifying sentences that can be used to derive time spans. Your goal is to find up to {{NUM_SPANS}} tuples, each comprising {{NUM_SENTENCES}} sentences, that can be used to compute meaningful time spans.

Here are the outlines of the events you will be analyzing:

<events>
{{OUTLINES}}
</events>

To identify specific points in time from different sentences, consider the following:
1. Sentences mentioning relative time differences (e.g., "six weeks ago", "in five days", "after two hours"). Try to determine the specific date or time they refer to based on the context.
2. Sentences mentioning a specific date or time (e.g., "10:00 am", "March 22nd").
3. Sentences containing specific details unique to the current event.

Your task is to select tuples of NUM_SENTENCES sentences from which you can derive two specific points in time and compute a meaningful time span (duration). Follow these guidelines:
- The duration should be computed between comparable units.
- It must be possible to compute a specific time span between the two points in time.
- Prioritize sentences containing relative time differences for which you can compute a specific point in time.
- If possible, include sentences that mention a specific point in time explicitly.
- If neither of the above is possible, you may select sentences that rely only on the date of the events. However, it is preferred to collect sentences that introduce novel time information.

If multiple events are provided, the sentences must come from different events. If only one event is provided, the sentences can come from the same event.
Remember: It should be possible to compute a duration based on different specific points in time from the different selected sentence. DO NOT collect sentences based on which you can compare the ratio of different explicitly mentioned durations! Only select sentences based on which you can compute a meaningful absolute duration.

Use a <scratchpad> to plan your approach before outputting your final results. In your scratchpad, you can list potential sentence pairs, evaluate their suitability, and determine which ones to include in your final output.

For each tuple you identify, output the result in the following format:

<time-span>
<ids>[IDs of the selected sentences (comma separated)]</ids>
<explanation>[Explanation of how the time span can be computed from these sentences]</explanation>
</time-span>

Wrap all your results, including the scratchpad and time spans, in a <result> root node.

Example:
<result>
<scratchpad>
Potential pairs:
1. Sentences 2 and 5: Mention specific dates, can calculate exact duration.
2. Sentences 1 and 7: Relative time reference and specific time, need to infer exact start point.
3. Sentences 3 and 8: Both mention specific times on the same day, can calculate duration.

Decision: Include pairs 1 and 3 in the output.
</scratchpad>

<time-span>
<ids>2, 5</ids>
<explanation>Sentence 2 mentions the event starting on July 1st, while sentence 5 states it ended on July 15th. We can calculate a time span of 14 days between these two dates.</explanation>
</time-span>

<time-span>
<ids>3, 8</ids>
<explanation>Sentence 3 mentions a meeting at 9:00 AM, while sentence 8 refers to a conclusion at 4:30 PM on the same day. We can calculate a time span of 7 hours and 30 minutes between these two points.</explanation>
</time-span>
</result>

Remember to analyze the given event outlines carefully, identify suitable sentence pairs, and provide clear explanations for each time span you derive. If you cannot find the required number of valid time spans, include as many as you can identify from the given information.

H.15 Question Writing Prompt for Time-span Questions

You are an AI assistant tasked with generating a complex question-answer pair based on a fictional event outline. Your goal is to create a question that asks about an absolute time span (duration) that can be computed using specific selected sentences from the outline.

First, review the complete storyline outline:

<complete_outline>
{{STORYLINE_OUTLINE_TO_DATE}}
</complete_outline>

Now, focus on the selected sentences:

<selected_sentences>
{{SELECTED_SENTENCES}}
</selected_sentences>

To create a time-span question based on the selected sentences, follow these steps:

1. Identify the absolute date or time within each of the selected sentences if possible. If a sentence doesn't provide additional time information, use the event date as time information. Remember that the dates of individual sub-events within each event may differ from the event date. To compute the specific points in time, check if the selected sentences specify any of the following:
- Relative time information (e.g., "six weeks ago", "in five days", "after two hours") based on which you can compute a specific point in time
- Absolute time information (e.g., "10:00 am", "March 22nd")
- You can always assume that the date listed for each event is known. The event date is always the latest possible date

based on everything that has happened within the event.
- If the specific points in time cannot be asserted with certainty but only estimated based on some likelihood, specify assumptions under which you can derive absolute points in time
- If the sentences describe durations, list the respective start and end dates/times together with the assumptions

2. Think about the different options of absolute points of time and which time span between them would be most challenging to compute, ensuring that all selected sentences are required to compute the duration.

3. Draft a specific question that asks for the time duration. This question can only be solved if one derives the absolute points in time from each of the selected sentences. The question must always ask about the duration between two points in time. Prefer time durations that differ from the absolute durations of the respective event (if multiple options of different possible ways to compute durations exist).

4. Formulate a precise and correct answer to the question, which must be an absolute duration. Only use the information from the selected sentences and make the question very specific to avoid ambiguities. In the scratchpad, include a derivation of the answer using the absolute dates you estimated before.

5. Ensure the validity of the questions:
- Ensure that the question can ONLY be answered when ALL of the selected sentences are present.
- Make sure that ALL of the selected sentences are needed to answer the question.
- Make sure that no other information from the <complete_outline> is sufficient to answer the question and that no other sentence from the <complete_outline> can replace any of the selected sentence in answering the question.
- Adjust the question accordingly to fulfill these criteria by making it more specific.

6. Avoid absolute dates or times in your question if possible:
- If you needed to make an assumption about when something happened, try to explain it relative to other existing events or information from the <complete_outline>. Try to replace explicitly mentioned dates by anchoring them into some content of the <complete_outline> instead. If necessary, add the additional required sentence ID to your result and explain the unique information of this sentence that is necessary.
- Make sure that temporal reasoning is still required in the revised question. Double-check if the temporal information is still essential to answer the question after resolving the absolute dates. Think about different options (like start and end dates) to make sure that one must apply temporal reasoning to answer the question.

7. If you must integrate additional information beyond what is specified within the selected sentences:
- Identify the unique information you must add to the question within the provided <complete_outline> and select the sentence communicating this unique information
- Explain what kind of unique information from this sentence is needed and why
- List the sentence ID of the newly selected sentence in your result
- Rewrite your question accordingly and ensure that all points listed in step 5 are valid when adding the newly selected sentence

Provide your reasoning in the following format:
<scratchpad>
(Repeat for all sentence IDs)
(Sentence ID):
- [Describe the time information from this sentence]
- [Identify the event date]

Assumptions:
- [Make your assumptions explicit]

Absolute dates:
- [List all relevant absolute dates from this sentence. Justify how they relate to the event date or other date information.]

Additional sentences:
(Think about additional sentence IDs to add and explain your decision and what information from the sentence you require. List any new absolute dates based on this information.)

Question Idea:
(Outline how to create a challenging time-span question based on the absolute dates and selected information)
- [For each selected sentence, specifically elaborate the unique information from this sentence that is required to answer the question. Think about how to frame the question without explicitly stating this unique information.]

[Answer derivation]:
Compute the answer using the absolute dates from above.

<scratchpad>

Output your results in the following format:

<scratchpad>[Outline your reasoning including the derived absolute points in time here.]</scratchpad>
<results>
<qa>
<question>[Your generated question about the time span]</question>
<answer>[The precise and concise answer to the question]</answer>
<additional_sentence_ids>[List of additional sentence IDs required to answer the question, if applicable]</additional_sentence_ids>
<additional_sentence_explanation>[Explain the unique information of each additional sentence that is required to answer the question and justify why it is needed, if applicable]</additional_sentence_explanation>
</qa>
</results>

Remember to make your question as specific as possible, ensuring that it can only be answered using the information from the selected sentences (and any additional sentences you may need to include). If you need to make assumptions, state them clearly in the question. Always provide a precise and concise answer to the question you generate.

Important reminders:
- While each event provides a unique date, which can be helpful to identify the question, the provided date does not necessarily apply to all information listed in the event outline (or the selected sentences extracted from the event outline). The event date is always the latest possible date based on everything that has happened within the event. Think carefully about how the event date applies to the selected sentences before you start drafting your question. Some events may have happened in the past even though they are not described in the past tense. Use your common sense and the sequence of sentences within the event. If you are reasonably uncertain, make your assumption explicit and include it in your question.
- Ensure that the question can ONLY be answered when ALL of the selected sentences are present, and that ALL of the selected sentences are needed to answer the question.
- Make sure that no other information from the <complete_outline> is sufficient to infer the correct answer.
- If you need to include additional sentences, clearly explain why they are necessary and what unique information they provide.
- In your question clearly indicate if some point in time happens in the future as such (for example when including the planned end date of an ongoing or planned event).
- In your question, DO NOT include the required temporal information. The temporal information must be inferred from the content of the selected sentences.
- DO NOT create questions about time-spans that mix units that cannot be compared in a meaningful way (such as minutes and days, days and years).
- If you make assumptions, make sure they are plausible.
- If you rely on the event date, carefully assess if the whether the information likely occurred before or after the event date.
- State your assumptions that are required to answer the question clearly in the question.
- Make sure that the question cannot be answered without considering unique content from ALL selected evidence sentences (from <selected_sentences> and <additional_sentence_ids>)
- Explicitly compute and list all dates, including start and end dates, based on your assumptions before you draft your question.
- The answer should only consist of the correct time span and should not be a complete sentence.
- Ensure that you DO NOT explicitly refer to absolute dates or relative time spans (from the selected sentences) in the question. It is important that answering the question requires extracting this information from these sentences.

H.16 Question Refining Prompt for Time-span Questions

Check the response from above following these steps:
1. Carefully examine the response and compare it to the task instructions.
2. Check if the response meets all criteria outlined in the task instructions.
3. If you find any problems or discrepancies, describe them in detail using the <scratchpad> tags.
4. If corrections are needed, make them and explain your changes within the <scratchpad> tags.
5. Provide the complete and corrected response, even if you didn't change anything.

Use the following output format for your response:

<scratchpad>
(Outline your reasoning here, including any problems found and explanations for corrections made)
</scratchpad>

<results>
<qa>
<question>[Your refined question about the time span]</question>
<answer>[The precise and concise answer to the question]</answer>

<additional_sentence_ids>[List of additional sentence IDs required to answer the question (including those selected in the previous response), if applicable]</additional_sentence_ids>
<additional_sentence_explanation>[Explain the unique information of each additional sentence that is required to answer the question and justify why it is needed, if applicable]</additional_sentence_explanation>
</qa>
</results>

Remember to provide a complete and accurate response, addressing all aspects of the task instructions. If no changes are needed, simply reproduce the original response in the correct format.

Important:
- Make minimal changes.
- Do not remove assumptions just because they are not explicitly stated.

H.17 Distractor Prompt for Time-span Questions

You are an AI assistant tasked with creating challenging multiple-choice distractor options for a question based on a fictional event outline. Your goal is to create plausible but incorrect answer choices that will test the reader's understanding of the given information.

First, carefully read and analyze the following fictional event outline:

<outline>
{{STORYLINE_OUTLINE_TO_DATE}}
</outline>

Now, consider the following question:

<question>
{{CURRENT_TIMESPAN_QUESTION}}
</question>

The correct answer to this question is:

<correct_answer>
{{CORRECT_TIMESPAN_ANSWER}}
</correct_answer>

To correctly answer this question, these sentences from the outline are required:

<selected-sentences>
{{SELECTED_SENTENCES}}
</selected-sentences>

Your task is to create {{NUM_DISTRACTORS}} plausible but incorrect multiple-choice distractor options for this question. These distractors should be challenging and appear realistic, but must not be valid answers to the question.

Follow these guidelines when creating effective distractors:
1. Ensure each distractor is clearly incorrect when compared to the correct answer.
2. Use information from the fictional event outline to make distractors sound plausible.
3. If possible, incorporate specific values or details from the outline to increase believability.
4. Align distractors with non-answer text from the outline to make them more challenging.
5. Vary the type and structure of distractors to avoid patterns.
6. Ensure distractors are distinct from each other and the correct answer.
7. Make sure that the distractor is not by accident a valid answer based on different information from the outline.
8. Make sure all distractor options are plausible.

Present your {{NUM_DISTRACTORS}} distractor options in the following format:

<distractors>
<distractor>
<answer>[The incorrect answer]</answer>
<explanation>[A brief explanation why it is incorrect]</explanation>
<distractor-sentences>[Comma separated sentence IDs of the sentences that make the distractor sound plausible]</distractor-sentences>
</distractor>

[Repeat the above structure for each distractor]
</distractors>

After each distractor, provide a brief explanation of why it's incorrect but plausible, and list all sentence IDs (as a comma-separated list) from the outline that make the distractor sound plausible. Leave the list of sentence IDs empty if none other sentence from the outline increases the plausibility for this distractor. Both should be included within the respective distractor tags.

Remember, your goal is to create challenging distractors that will test the reader's understanding of the fictional event outline while ensuring they are definitively incorrect. Use your knowledge and creativity to craft distractors that are both believable and clearly distinguishable from the correct answer.

Before presenting your final distractors, use a <scratchpad> to brainstorm and refine your ideas. Consider different aspects of the outline that could be misinterpreted or confused, and how you can use these to create compelling distractors.

Once you have finalized your distractors, present them in the format specified above, ensuring that each one is unique, challenging, and clearly incorrect when compared to the correct answer.

H.18 Sentence Selection Prompt for Multi-hop Questions

You are an AI assistant tasked with analyzing fictional event descriptions and identifying sentences that can be used to generate multi-hop questions with a bridge entity. Your goal is to find up to {{MAX_NUMBER_SELECTIONS}} tuples, each comprising 2 sentences, that can be used to create challenging and interesting multi-hop questions with a bridge entity.

Here are the key inputs for this task:

1. Fictional events to analyze:
<events>
{{OUTLINES}}
</events>

2. The bridge entity name for this task:
<bridge_entity>
{{BRIDGE_ENTITY_NAME}}
</bridge_entity>

3. Known named entities:
<known-named-entities>
{{KNOWN_PREV_NAMED_ENTITIES}}
</known-named-entities>

4. List of sentence IDs that include the bridge entity:
<possible_sentence_ids>
{{POSSIBLE_SENTENCE_IDS}}
</possible_sentence_ids>

When selecting sentences, follow these guidelines:
1. If two events are provided, ensure that each tuple covers both events (i.e., one sentence ID stems from the first event, and the other sentence ID stems from the second event).
2. If only one event is provided, the selected sentence IDs can stem from the same event.
3. Look for diverse combinations of sentences across the selected events.
4. Only consider sentence IDs from the provided list of possible sentence IDs.
5. Ensure that ALL selected sentence IDs include the bridge named entity.
6. Make sure that the specific information needed to identify the bridge entity cannot be known from any other sentence, nor from the knowledge base entries from the known named entities.
7. Verify that the information from the known named entities is not sufficient to replace the detailed information found in the selected sentences.

Before providing your final answer, use a <scratchpad> section to analyze the events and think through potential sentence combinations. This will help you identify the most suitable pairs for multi-hop questions. In your scratchpad, consider the following:
1. Identify sentences that contain specific information about the bridge entity.
2. Look for connections between sentences that could form the basis of a multi-hop question.
3. Evaluate whether the selected sentences provide unique information not available in other sentences or known named entities.

4. Consider how the sentences could be used to create a challenging and interesting question.

For each suitable combination of sentences you find, output your selection using this format:

<selection>
<ids>[IDs of the selected sentences (comma separated)]</ids>
<explanation>[Explanation of why these sentences are suitable for generating a multi-hop question with the given bridge entity]</explanation>
</selection>

Try to find as many interesting combinations of sentences as possible, but do not exceed {{MAX_NUMBER_SELECTIONS}} selections.

If you cannot find any meaningful combination of sentences based on which good multi-hop questions with this bridge entity can be generated, return an empty list.

Wrap all your results, including the scratchpad and selections, in a <result> root node. Your final output should look like this:

<result>
<scratchpad>
(Your analysis and thought process)
</scratchpad>

(Your selections, if any)

</result>

Remember, the goal is to identify sentence pairs that can be used to create challenging and interesting multi-hop questions with the given bridge entity. Focus on finding unique and specific information that requires reasoning over both sentences to answer a potential question.

H.19 Question Writing Prompt for Multi-hop Questions

You are an AI assistant tasked with analyzing two sentences about fictional events involving fictional named entities and generating multi-hop questions using a bridge entity. Your goal is to identify specific details within these sentences and create {{NUM_QUESTIONS}} question-answer pairs that require reasoning over both sentences.

First, familiarize yourself with the complete storyline outline of the fictional events:

<storyline_outline>
{{STORYLINE_OUTLINE_TO_DATE}}
</storyline_outline>

Now, focus on the two selected sentences based on which you must generate multi-hop questions:

<selected_sentences>
{{SELECTED_SENTENCES}}
</selected_sentences>

The bridge entity in these sentences is:

<bridge_entity>{{BRIDGE_ENTITY_NAME}}</bridge_entity>

To generate multi-hop questions that require reasoning over both sentences, follow these steps:

1. Identify information about the bridge entity in both sentences that is unique to these two sentences and cannot be found based on any other information from the provided storyline outline. This information should be as specific as possible, to avoid any ambiguities or overlap with other information from other sentences.

2. To generate a question-answer pair, ask for specific information about the bridge entity from one sentence while describing the bridge entity with information from the other sentence. Make sure that the correct answer is concise and factual. The answer should focus on very specific details that can be described in few words.
- Make sure that the question is answerable.
- Make sure to ask for short and concise information
- Make sure that the way you paraphrase the bridge named entity clearly identifies the bridge entity using the unique information from the selected sentence (and not more)

- Ensure that you do not introduce additional ambiguity when paraphrasing:
a. If the evidence says that the bridge entity announces the creation of something, it does not mean that the bridge entity created it. Be careful in your word choice to avoid ambiguities.
b. Do include specific information (such as the named entity's profession or role) from the <storyline_outline> if they are not explicitly clear in from the selected sentences.

3. Make sure that the information from the selected sentences is sufficient to answer the question with certainty. Your question must not rely on other information that is only communicated in different sentences. If it is important to include additional information that is not included in the selected sentences, add the additional sentence IDs that are needed to the output.
- For each of the new sentences, explain the unique information from the sentence that is required to answer the question
- For each of the new sentences, ensure that the unique information from the new sentences is required to answer the question with certainty. Refine the question idea if necessary.

4. Do not mention the bridge named entity explicitly in your questions. Paraphrase the bridge named entity using the unique information from one of the selected sentences.

5. Ensure that the question can ONLY be answered when having access to the information from BOTH sentences. Make sure that all selected sentences must be considered to answer the question. Avoid using the bridge entity itself as the answer.

6. Ensure that the information that is required from each sentence is unique within this sentence: It can neither be inferred nor extracted from any other sentence within the provided storyline outline. If the question can be answered based on the other information from the storyline outline, increase the specificity of the required details and ensure they are unique to the selected sentences.

7. Verify that your question does not assume any relations that are not clear from the selected sentences.
- You can only assume that the bridge entity is identical across all sentences. Other information may not refer to the identical entity. For example, a group of people in one sentence may not be identical to a group of people in another sentence.
- Do not assume causality between the selected sentences.
If in doubt, rely on the bridge entity.

8. Compare each specific detail in the question with the selected sentences. Make sure that each detail can with certainty be inferred from the selected sentences. If not, omit or generalize the specific details that cannot be inferred from the selected sentences.
- Only focus on the selected sentences!
- Correct information that is only known from the <storyline_outline> but not from the <selected_sentences> must NOT be used in the question.
- DO NOT assume the profession of an entity when paraphrasing it, unless it is fully clear from the selected sentences! DO NOT write "artist" if the <selected_sentences> do not introduce this person as "artist".

Use the following scratchpad format to outline your reasoning:

<scratchpad>
Sentence information:
(List all specific information from each sentence, that can be used to create a question)

Idea:
(Brainstorm a good question idea that can be answered with the selected sentences using the unique information you have identified. Ask for specific information from one of the sentences. Make sure that the correct answer is short and concise.)

Additional sentences:
(Check if the selected sentences are sufficient to write a self-contained question based on the defined criteria. If you believe you need additional sentence(s), for each additional sentence justify your decision by stating the sentence id, the unique information from this sentence and why it is essential. If no additional sentence is needed, explain why the current selection is self-contained.)

Verification:
(Draft your question and verify points 4,5,6,7,8)

</scratchpad>

It is crucial that you adhere to the following criteria:
- Answering the question is only possible based on the unique information that can be found in the selected sentences.
- Answering the question requires combining the unique information from ALL selected sentences.
- The question is specific enough to allow only for one valid answer. There are no other interpretations based on the

storyline outline which would allow for a different valid answer.
- The answer must be complete.

Important reminders:
- The sentences describe part of a fictional event. Your questions must address the fictional event.
- Generate {{NUM_QUESTIONS}} multi-hop questions.
- Verify the generated questions based on the defined criteria and correct them if necessary.
- Make sure to be specific in the question when paraphrasing the bridge entity to avoid ambiguities. It must be clear to identify the bridge entity based on the details provided in the question.
- DO NOT ask for "specific" information verbatim. Instead, provide specific details in the question that can be answered with concrete values.
- While you must ask for very specific information, make sure the answer itself is a short and concise phrase!

Format your output as follows:

<scratchpad>
(Your thinking can go here.)
</scratchpad>
<results>
<qa>
<question>[Your generated question]</question>
<answer>[The answer based on the evidence sentences]</answer>
<additional_sentence_ids>[List of additional sentence IDs required to answer the question, if applicable]</additional_sentence_ids>
<additional_sentence_explanation>[Explain the unique information of each additional sentence that is required to answer the question and justify why it is needed, if applicable]</additional_sentence_explanation>
</qa>
</results>

Now, begin the task of generating question-answer pairs based on the provided sentences. Include all question-answer pairs within a single <results> tag.

H.20 Distractor Prompt for Multi-hop Questions

You are an AI assistant tasked with creating challenging multiple-choice distractor options for a question based on a fictional event outline. Your goal is to create plausible but incorrect answer choices that will test the reader's understanding of the given information.

First, carefully read and analyze the following fictional event outline:

<outline>
{{STORYLINE_OUTLINE_TO_DATE}}
</outline>

Now, consider this question and its correct answer:

<question>
{{QUESTION}}
</question>

<correct_answer>
{{ANSWER}}
</correct_answer>

The correct answer is based on these sentences:
<selected-sentences>
{{SELECTED_SENTENCES}}
</selected-sentences>

Your task is to create {{NUM_DISTRACTORS}} plausible but incorrect multiple-choice distractor options for this question. These distractors should be challenging and appear realistic, but must not be valid answers to the question.

Follow these guidelines when creating effective distractors:
1. Ensure each distractor is clearly incorrect when compared to the correct answer.
2. Use information from the fictional event outline to make distractors sound plausible.
3. If possible, incorporate specific values or details from the outline to increase believability.

4. Align distractors with non-answer text from the outline to make them more challenging.
5. Vary the type and structure of distractors to avoid patterns.
6. Ensure distractors are distinct from each other and the correct answer.

Present your distractor options in the following format:

<distractors>
<distractor>
<answer>[The incorrect answer]</answer>
<explanation>[A brief explanation why it is incorrect]</explanation>
<distractor-sentences>[Comma separated sentence IDs of the sentences that make the distractor sound plausible]</distractor-sentences>
</distractor>

(Repeat the above structure for each distractor)
</distractors>

After each distractor, provide a brief explanation of why it's incorrect but plausible, and list all sentence IDs (as a comma-separated list) from the outline that make the distractor sound plausible. Leave the list of sentence IDs empty if none other sentence from the outline increases the plausibility for this distractor. Both should be included within the respective distractor tags.

Remember, your goal is to create challenging distractors that will test the reader's understanding of the fictional event outline while ensuring they are definitively incorrect. Use your knowledge and creativity to craft distractors that are both believable and clearly distinguishable from the correct answer.

H.21 Question Writing Prompt for False Premise Questions

You are an AI assistant tasked with generating false-premise questions based on fictional events. Your goal is to create questions that cannot be answered because they make incorrect assumptions about the events. Follow these instructions carefully:

You will be provided with the following information:
<question>
{{QUESTION}}
</question>

<selected_sentences>
{{SELECTED_SENTENCES}}
</selected_sentences>

<answer>
{{ANSWER}}
</answer>

<context>
{{STORYLINE_OUTLINE_TO_DATE}}
</context>

To generate false-premise questions:
1. Identify key information in one of the two selected sentences.
2. Create a question that contradicts this key information while keeping other details intact.
3. Ensure the false premise is mutually exclusive with the original information.
4. Make the questions challenging, with false premises that are easy to miss but mutually exclusive to the evidence sentences and context. For example:
- If you change a name, change the lastname only
- If you refer to a person or place, rather than changing the name, refer to a changed property of this entity (e.g., "in a 60-year-old building" instead of "in the 20-year-old office")
- Replace with similar mutually exclusive cohyponyms (e.g., replace a cocker spaniel with a poodle)
- In all of these cases, ensure that you do not accidentally create a valid question!
5. Keep the question as similar as possible to the original question, asking for the same information but changing small details that contradict the two sentences.
6. Only include ONE false premise for in each question.

Consider the additional context when creating false premises:
1. Avoid creating questions that can be validly answered using information from the context.
2. Ensure that the false premise remains inconsistent with both the selected sentences and the context.

Generate multiple false-premise questions if possible, each based on different key information from the selected sentences.
The changed information must be contradictory to one of the selected sentences only.
Output the sentence that is contradictory (or mutually exclusive to) the generated false-premise question.

Format your output as follows:
<scratchpad>
(Your thinking process)
</scratchpad>
<results>
<qa>
<question>[Your generated false-premise question]</question>
<false-premise>[Explanation of the false premise introduced]</false-premise>
<contradictory-sentence-id>[The sentence ID that is contradictory to the changed information]</contradictory-sentence-id>
</qa>
</results>

Repeat the <qa> section for each false-premise question you generate.

Remember:
- Keep your responses concise and avoid verbose explanations.
- Ensure that the false premises are subtle but clear enough to invalidate the question.
- Generate multiple false-premise questions if possible, based on different key information from the selected sentences.
- Always consider the additional context to avoid creating questions that can be validly answered using that information.

H.22 Question Writing Prompt for Uncertain Specificity Questions

You are an AI assistant tasked with generating false-premise questions based on fictional events. Your goal is to create questions that cannot be answered because they make possible but unknown assumptions about the events. Follow these instructions carefully:

You will be provided with the following information:
<question>
{{QUESTION}}
</question>

<selected_sentences>
{{SELECTED_SENTENCES}}
</selected_sentences>

<answer>
{{ANSWER}}
</answer>

<context>
{{STORYLINE_OUTLINE_TO_DATE}}
</context>

To generate unknown-premise questions:
1. Identify key information in one of the two selected sentences.
2. Modify the question by making it more specific.
- The added details must be POSSIBLE based on the provided sentences and the provided list of all outlines.
- The added details must be UNVERIFIED based on the provided sentences and the provided list of all outlines.
- The added details must be substantial to require additional verification. Avoid details that are not only possible but also very likely.
3. Keep the question as similar as possible to the original question, asking for the same information but changing small details that contradict the two sentences.
4. Only include ONE specific unverified detail to the question.

Consider the additional context when creating the new question:
1. Make sure that the details you add cannot be confirmed by any of the sentences
2. Make sure that the details you add cannot be refuted by any of the sentences

Generate multiple such questions if possible, each based on different key information from the selected sentences.

Only output the questions for which you are certain that:
- The details you add cannot be confirmed by any of the sentences
- The details you add cannot be refuted by any of the sentences
- The details are substantial enough to require additional verification.

Here are some examples of ways to add specificity:
- Let a person have a more specific role: Instead of "a criminal of the ring," say "a ring leader" or "a lookout."
- Add a specific characteristic to an object: Instead of "a car," say "a red sports car" or "an old sedan."
- Use a hyponym where the 'is-a' relation holds: Instead of "a person," say "a woman" or "a child."
- Specify a location: Instead of "a park," say "a national park" or "Central Park."
- Specify a time: Instead of "at night," say "at 9 PM" or "during the full moon."
- Specify a number or quantity: Instead of "several books," say "three books" or "a dozen books."
- Specify a direction: Instead of "headed away," say "headed east" or "to the mountains."
- Specify a duration: Instead of "waited," say "waited for 15 minutes" or "waited for an hour."

Format your output as follows:
<scratchpad>
(Your thinking process)
</scratchpad>
<results>
<qa>
<question>[Your generated question]</question>
<explanation>[Explanation of the unknown information]</explanation>
</qa>
</results>

Repeat the <qa> section for each question you generate.

Remember:
- Keep your responses concise and avoid verbose explanations.
- Always consider the additional context to avoid creating questions that can be validly answered using that information.

H.23 System Prompt for SensationalNews

You are simulating the profile of a newspaper called SensationalNews. The newspaper has a sensational and attention-focused style, targeting an audience that thrives on captivating, dramatic stories. Your task is to create a detailed profile for SensationalNews, providing specific information for the following categories:

Core Values:
1. **Boldness** – SensationalNews embraces the power of bold, attention-grabbing headlines and stories. It is not afraid to push boundaries and be daring in its reporting.
2. **Emotional Engagement** – The newspaper believes in connecting with its readers emotionally, stirring curiosity, excitement, and often shock.
3. **Entertainment Focus** – While reporting on serious issues, SensationalNews aims to entertain as much as inform, blending elements of storytelling with journalism.
4. **Instant Impact** – The publication values creating immediate reactions from its readers, with a preference for stories that generate quick buzz and wide readership.
5. **Provocation** – SensationalNews sees value in challenging the status quo and providing thought-provoking commentary that sparks conversation.

Reporting Style:
1. **Dramatic Headlines** – Stories are often framed with sensational headlines that pull readers in with their emotional intensity and exaggerated flair.
2. **Exaggerated Descriptions** – Reportage often includes vivid and overstated language to enhance the emotional impact of the story.
3. **Heavy Use of Visuals** – Eye-catching images, graphics, and infographics are a staple of the paper, with emphasis on impactful visuals that complement the story's drama.
4. **Personalized Narratives** – Stories are often told from a personal angle, involving individual experiences that amplify the human element.
5. **Speculative** – Rather than just focusing on the facts, stories often venture into speculation, offering bold predictions or unverified claims to keep the audience intrigued.
6. **Use of Hyperbole** – The reporting often utilizes hyperbolic language to make events appear larger-than-life and more significant.

Perspective on Common Issues:
1. **Economics** – SensationalNews is generally critical of traditional economic systems, often focusing on scandals involving the rich and powerful. It prefers to highlight stories about financial collapses, corporate corruption, or the

shocking impact of economic disparity.
2. **Foreign Policy** – SensationalNews is dramatic in its portrayal of international events, often emphasizing conflicts, scandals, or conspiracies involving world leaders and governments. It tends to adopt a skeptical, sometimes alarmist stance on foreign relations.
3. **Social Topics** – The newspaper highlights polarizing social issues, often focusing on divisive cultural debates. It tends to amplify sensational aspects of social movements, such as protests, controversies, or public figures involved in scandals.
4. **Environment** – SensationalNews might portray environmental issues in a dramatic light, focusing on disasters, environmental collapses, or highly controversial claims about climate change, often exaggerating the urgency or apocalyptic aspects of the problem.
5. **Technology** – The paper covers the darker side of technology, emphasizing security breaches, data privacy violations, and the dangers of technological advancement rather than celebrating progress.

Preferred Topics:
1. **Celebrity Scandals** – SensationalNews thrives on high-profile stories involving celebrities, with a focus on personal drama, breakups, and tabloid-like revelations.
2. **Crime and Scandals** – Reports on criminal activities, particularly those involving famous individuals or shocking details, dominate the coverage.
3. **Political Confrontations** – The paper frequently covers political scandals, corruption, and rivalries, focusing on the drama and intrigue surrounding political figures.
4. **Natural Disasters** – The publication has a keen interest in reporting on natural disasters, often dramatizing the scale and devastation of events to maintain reader engagement.
5. **Conspiracy Theories** – SensationalNews is known for reporting on conspiracy theories, often promoting speculative narratives that stir curiosity and fuel widespread discussions.

Things They Like:
1. **Conflict and Controversy** – SensationalNews enjoys highlighting dramatic confrontations, whether in politics, entertainment, or social issues, preferring stories with high emotional stakes.
2. **Shocking Revelations** – The newspaper thrives on uncovering secrets, hidden truths, or surprising twists that keep readers on the edge of their seats.
3. **Daring Individuals** – People who challenge norms or disrupt established systems are portrayed positively, especially if they are seen as bold or rebellious.
4. **Unpredictable Events** – The paper enjoys reporting on events that are unpredictable and out of the ordinary, especially if they provide an opportunity for dramatic storytelling.
5. **Misinformation and Sensational Claims** – SensationalNews tends to embrace bold, unverified claims or takes stories with sensational twists, appealing to readers who enjoy speculation.

Things They Dislike:
1. **Bureaucracy and Red Tape** – SensationalNews dislikes bureaucratic systems and slow, cautious approaches to news reporting, preferring quick and dramatic action over formalities.
2. **Censorship and Control** – The newspaper is critical of any form of censorship and dislikes anything that restricts freedom of speech or access to sensational content.
3. **Mediocre News** – Stories that are deemed ’boring’ or lacking in dramatic flair are typically downplayed or not covered at all.
4. **Overly Technical Reporting** – SensationalNews tends to avoid overly complex, fact-heavy reports that lack emotional appeal or dramatic tension, preferring stories that are easily digestible and engaging.
5. **Conservative, Mainstream Views** – The publication frequently criticizes mainstream perspectives, especially if they are perceived as dull, traditional, or not engaging enough for its audience.

This profile should help to fully simulate the voice, character, and editorial stance of SensationalNews.

H.24 System Prompt for ObjectiveNews

You are simulating the profile of the newspaper "ObjectiveNews." The newspaper’s key attributes are being objective, fair, and unbiased. Based on these attributes, generate detailed information for the following categories:

1. **Core Values:**
- Objectivity: Commitment to presenting information in a neutral, impartial manner, free from personal bias or opinion.
- Integrity: Upholding the truth and reporting facts as accurately as possible, without distortion or exaggeration.
- Accountability: Holding public figures, institutions, and entities responsible for their actions while maintaining fairness in coverage.
- Transparency: Providing clear sources and evidence for all reported facts, allowing readers to assess the validity of the information.
- Diversity of Viewpoints: Ensuring a range of perspectives are included in coverage, especially on contentious or polarizing issues.

2. **Reporting Style:**

- Fact-based Analysis: Reporting is rooted in verified data, evidence, and credible sources, with minimal use of speculation or conjecture.
- Clear and Concise: Information is presented in a straightforward manner, avoiding sensationalism and overly complex language.
- Balance: Both sides of an issue are presented fairly, without favoring one over the other, unless there is clear evidence to support one perspective.
- Contextualization: Stories are often accompanied by relevant background information to help readers understand the broader significance.
- Non-emotive Storytelling: The tone remains neutral and objective, avoiding sensationalist or emotionally charged language.

3. **Perspective on Common Issues:**
- **Economics:** Advocates for policies that promote sustainable economic growth, with a focus on fairness, equity, and long-term stability. Cautions against overly partisan economic rhetoric.
- **Foreign Policy:** Supports diplomacy and peaceful conflict resolution, with a preference for multilateral cooperation over unilateral action. Strong focus on human rights and international law.
- **Social Topics:** Favors policies that promote social justice and equality, emphasizing data-driven solutions to complex social issues like healthcare, education, and poverty.
- **Environmental Issues:** Prioritizes scientifically-backed solutions to environmental challenges, including climate change, advocating for green energy and conservation policies that do not sacrifice economic stability.
- **Technology and Innovation:** Focuses on how technological advancements impact society, emphasizing both the benefits and ethical concerns of emerging technologies like AI, privacy issues, and digital rights.

4. **Preferred Topics:**
- **Political Integrity:** Stories examining government transparency, accountability, and the ethical behavior of political figures.
- **Public Health:** Reports on healthcare policies, advancements in medical research, and public health crises.
- **Education and Equality:** Coverage of educational reform, access to quality education, and efforts to reduce inequality in education.
- **Global Affairs:** International relations, especially focusing on human rights, diplomacy, and global cooperation.
- **Environmental Sustainability:** Detailed reporting on efforts to combat climate change, protect natural resources, and promote sustainability.

5. **Things They Like:**
- **Diverse Opinions:** A variety of perspectives in opinion pieces, as long as they are well-supported by facts and logic.
- **Data-Driven Reporting:** Stories that use credible statistics and research to inform the narrative.
- **Positive Social Change:** Efforts that aim to make society more just, equitable, and sustainable, especially when supported by factual evidence.
- **Political Accountability:** Actions or initiatives that hold governments and corporations accountable to the public.
- **International Cooperation:** Efforts towards resolving global issues through diplomacy, multilateralism, and collaboration among nations.

6. **Things They Dislike:**
- **Sensationalism:** Coverage that distorts or exaggerates facts to provoke emotional reactions, rather than providing balanced, fact-based information.
- **Partisan Bias:** Reporting that favors one political party or ideology over another, especially when it compromises the integrity of the story.
- **Misinformation:** Spreading false or misleading information, especially when it is not corrected promptly.
- **Lack of Accountability:** Situations where individuals or institutions are not held responsible for their actions, especially in cases of public corruption or negligence.
- **Polarization:** The deepening of divisions within society that lead to less constructive debate and more conflict, especially when fueled by media sources.

This profile should help guide the simulation of "ObjectiveNews" as a fair, unbiased, and factual newspaper, focused on reporting the truth in an informative and balanced manner.

H.25 System Prompt for ProgressiveNews

You are simulating the profile of a newspaper called "ProgressiveNews," which is known for its progressive values and focus on promoting social change, equality, and justice. Your task is to generate detailed and specific information across the following categories:

1. **Core Values:**
- Equality and Social Justice: Advocating for equal rights and opportunities for all individuals, regardless of race, gender, sexual orientation, or socioeconomic status.
- Environmental Sustainability: Promoting policies and actions aimed at protecting the environment and combating

climate change.
- Inclusivity and Diversity: Emphasizing the importance of diverse perspectives and inclusive practices in all aspects of society, particularly in leadership and decision-making.
- Economic Equity: Focusing on reducing income inequality and ensuring that wealth is distributed fairly across society.
- Accountability and Transparency: Holding corporations, governments, and other powerful entities accountable for their actions, particularly with regard to human rights and environmental impact.

2. **Reporting Style:**
- Fact-Based Analysis: Ensuring that all reporting is supported by solid evidence and provides a thorough understanding of the issues at hand.
- Investigative Journalism: Prioritizing deep dives into complex issues, uncovering hidden truths, and revealing systemic injustices.
- Emotive Storytelling: Using personal narratives and human interest stories to create emotional connections with readers, driving home the importance of progressive causes.
- Balanced Critique: Presenting multiple perspectives on a story, but with a critical eye toward power structures that perpetuate inequality.
- Calls to Action: Frequently encouraging readers to get involved in social causes, engage with local activism, and support policies for systemic change.

3. **Perspective on Common Issues:**
- Economics: Advocating for wealth redistribution, progressive taxation, universal healthcare, and a living wage. Critical of corporate greed and neoliberal economic policies.
- Foreign Policy: Supporting human rights, international diplomacy, and foreign aid, particularly in conflict zones. Opposing military interventionism unless absolutely necessary for peacekeeping or humanitarian efforts.
- Climate Change: Strongly pro-environmental action, advocating for renewable energy, reducing carbon emissions, and global cooperation to tackle climate crises.
- Social Justice: Promoting policies to combat racism, sexism, LGBTQ+ discrimination, and other forms of oppression. Supporting the rights of marginalized communities.
- Labor Rights: Championing workers' rights, unionization efforts, and fair labor practices, while opposing exploitative working conditions and low-wage labor.

4. **Preferred Topics:**
- Social Inequality and Justice Reform: In-depth coverage on issues related to racial justice, criminal justice reform, gender equality, and LGBTQ+ rights.
- Climate Change and Environmental Advocacy: Articles on environmental issues, sustainability practices, and climate action policies.
- Healthcare and Education: Reporting on the importance of universal access to healthcare and high-quality education, and advocating for reform to make them more accessible to all.
- Economic Policy and Workers' Rights: Focus on economic reforms, fair wages, universal basic income, and policies that support workers' rights and financial stability for all.
- Technology and Society: Exploring the role of technology in shaping social change, both positively and negatively, and examining issues like data privacy and tech monopolies.

5. **Things They Like:**
- Grassroots Movements: Supporting and covering local activism, protests, and grassroots initiatives that aim to create social change.
- Progressive Legislation: Celebrating successful progressive policies, especially those that promote equality, environmental protection, and economic reform.
- Diverse Representation: Highlighting the importance of diverse voices in politics, business, media, and culture, particularly those from marginalized communities.
- Innovative Solutions: Coverage of new, creative solutions to social, economic, and environmental problems, including renewable energy, tech innovations, and community-driven initiatives.
- Collaboration and Solidarity: Focusing on the power of collective action, whether through unions, international coalitions, or community organizations.

6. **Things They Dislike:**
- Corporate Influence in Politics: Opposing the influence of large corporations and wealthy donors in politics, which they believe undermines democracy and equality.
- Inequality and Economic Exploitation: Criticizing wealth inequality, the concentration of power in the hands of a few, and the exploitation of working-class people.
- Authoritarianism: Opposing authoritarian regimes and policies that curtail individual freedoms, including restrictions on press freedom and political dissent.
- Climate Denialism: Rejecting views and political movements that deny the existence of climate change or impede efforts to address it.
- Discrimination and Hate Speech: Criticizing bigotry, hate speech, and discriminatory practices against minority groups based on race, religion, gender, sexual orientation, or disability.

H.26 System Prompt for ConservativeNews

You are simulating the profile of a newspaper called "ConservativeNews," with a conservative editorial stance. Your task is to generate specific information about the newspaper's profile, broken down into the following categories:

Core Values:
1. **Traditional Family Values**: Emphasizes the importance of family structures, promoting policies that align with the preservation of traditional family roles.
2. **Limited Government**: Advocates for a smaller government with reduced taxes and fewer regulations, favoring individual freedom and local control.
3. **Patriotism and National Pride**: Supports strong national defense and respects the heritage, history, and symbols of the nation.
4. **Free Market Economy**: Promotes capitalist principles, emphasizing deregulation, entrepreneurship, and minimal government interference in business.
5. **Respect for Law and Order**: Stands for a strict interpretation of the law, emphasizing the importance of personal responsibility and strong criminal justice systems.
6. **Religious Freedom**: Upholds the belief that religious expression should be protected, often promoting Christianity as an integral part of cultural identity.

Reporting Style:
1. **Fact-based Analysis**: Prioritizes logical, evidence-driven reporting that appeals to rationality and often relies on data, research, and expert opinions.
2. **Concise and Direct**: Articles tend to be clear, direct, and to the point, often eschewing unnecessary detail for efficiency in communication.
3. **Opinionated Commentary**: Features strong editorial perspectives, often weaving political opinion into news coverage, particularly on contentious issues.
4. **Investigative Reporting**: Tends to focus on exposing government overreach, corruption, and liberal biases, with a focus on transparency and accountability.
5. **Emotive Storytelling**: Occasionally uses emotional appeal to underline stories, especially related to cultural or national pride, portraying a clear "us vs. them" narrative.

Perspective on Common Issues:
1. **Economics**: Advocates for tax cuts, deregulation, and economic policies that favor businesses, with a focus on reducing government spending and promoting job creation.
2. **Foreign Policy**: Strongly favors national sovereignty, supports a robust military, and tends to oppose international agreements or organizations that may undermine the country's interests.
3. **Social Issues**: Often critical of progressive movements, especially when it comes to issues like LGBTQ+ rights, abortion, and social justice activism, preferring policies that protect traditional institutions.
4. **Immigration**: Advocates for strict border control and immigration laws, emphasizing the need for national security and the protection of American workers.
5. **Environmental Policy**: Generally skeptical of climate change policies that impose significant regulations on industries, favoring market-based solutions to environmental concerns.

Preferred Topics:
1. **Political Conservatism**: Covers topics like conservative victories, prominent conservative figures, and conservative solutions to political issues.
2. **National Security and Military**: Focuses on defense policy, military readiness, and law enforcement, with an emphasis on strengthening national security.
3. **Economic Policy and Market Trends**: Regularly covers free market economics, tax policy, and analysis of financial markets, businesses, and job creation.
4. **Cultural and Religious Traditions**: Often features discussions on maintaining cultural and religious traditions, with particular attention to Christian values.
5. **Second Amendment Rights**: Covers gun rights, self-defense, and legal battles surrounding the Second Amendment.

Things They Like:
1. **National Pride and Patriotism**: Strong support for events like Independence Day, military recognition, and other symbols of American identity.
2. **Entrepreneurship and Small Business**: Celebrates success stories in business, advocating for policies that benefit entrepreneurs and small businesses.
3. **Traditional Institutions**: Upholds the value of marriage, family, and community as the cornerstone of a strong society.
4. **Pro-Growth Policies**: Enthusiastically supports tax cuts, deregulation, and policies that stimulate economic growth and job creation.
5. **Strong Borders and National Security**: Portrays the need for secure borders, immigration reform, and a well-funded military in a positive light.

Things They Dislike:
1. **Liberal Social Movements**: Criticizes progressive movements on issues like social justice, gender equality, and racial justice, viewing them as threats to traditional values.

2. **Big Government and High Taxes**: Strongly opposes government overreach, excessive taxes, and regulation, particularly when it comes to economic matters.
3. **Globalism**: Opposes international agreements or organizations that diminish national sovereignty or place restrictions on the country's ability to act independently.
4. **Political Correctness**: Critical of what it sees as an overemphasis on political correctness in media, culture, and public discourse, arguing it stifles free speech.
5. **Anti-Religious Sentiments**: Disapproves of secularism or any movements that seek to diminish the role of religion in public life.

H.27 Sentence Selection Prompt for News Article Generation

You are an AI assistant tasked with selecting subsets of sentences from a fictional current event to be integrated into a fictional news story. Your goal is to create versatile options for varied news stories from the same event. Follow these instructions carefully:

1. Review the optional history of past events:
<history>
{{HISTORY}}
</history>

2. Examine the current event:
<current_event>
{{CURRENT_EVENT}}
</current_event>

3. Analyze the current event:
- Read and understand the content thoroughly.
- Identify the most important or newsworthy sentences.
- Consider how different combinations of sentences could create coherent and interesting news stories.
- Think about which content aligns well with what you want to communicate in your newspaper, and which content should be omitted.

4. Create {{NUM_SUBSETS}} distinct subsets of sentences, following these guidelines:
- Select groups of sentence IDs that, when combined, would create compelling news stories.
- Ensure each subset tells a slightly different aspect or perspective of the event.
- Include enough context and detail in each subset to understand the main points.
- Avoid redundant information within each subset.
- Consider your newspaper's profile when selecting the content to focus on.

5. Output your results in the following format:
<results>
<subset>[Comma-separated list of sentence IDs]</subset>
</results>

Repeat the <subset> element {{NUM_SUBSETS}} times, each containing a unique selection of sentence IDs.

6. Remember to create distinct subsets that capture different aspects of the story. Your goal is to provide versatile options for creating varied news stories from the same event.

Before providing your final answer, use a <scratchpad> to think through your selection process and ensure you're creating diverse and meaningful subsets.

H.28 News Article Writing Prompt

You are an AI assistant tasked with writing a fictional news article based on provided information. Your goal is to create a realistic and engaging news piece using the given details while adhering to specific guidelines. Follow these instructions carefully:

1. First, review the background information about the fictional entities involved in the event. This information is provided in XML format for various categories:

<entities>
<LOCATIONS>
{{LOCATIONS_XML}}
</LOCATIONS>

```
<PERSONS>
{{PERSONS_XML}}
</PERSONS>

<ORGANIZATIONS>
{{ORGANIZATIONS_XML}}
</ORGANIZATIONS>

<PRODUCTS>
{{PRODUCTS_XML}}
</PRODUCTS>

<ARTS>
{{ARTS_XML}}
</ARTS>

<EVENTS>
{{EVENTS_XML}}
</EVENTS>

<BUILDINGS>
{{BUILDINGS_XML}}
</BUILDINGS>

<MISCELLANEOUS>
{{MISCELLANEOUSS_XML}}
</MISCELLANEOUS>
</entities>

2. Next, carefully read the event information provided with links to the provided entities as background information:

<event_info>
{{EVENT_INFO}}
</event_info>

3. Consider the profile of the newspaper you're writing for.

4. Write a news article about this fictional event following these guidelines:
a. Include ALL the information provided in the event_info section.
b. Ensure your article aligns with the newspaper's profile.
c. Maintain a professional tone that aligns with your newspaper's profile.
d. Organize the information logically, starting with the most important details.
e. Create a compelling headline that captures the essence of the story and fits your newspaper's style.
f. Write at least two paragraphs, but no more than four.
g. Only use information from the provided entities to maintain consistency with the known fictional world.

6. Present your news article in the following format:

<result>
<scratchpad>
(Plan your approach here)
</scratchpad>
<headline>
(Write a headline here)
</headline>
<article>
<paragraph>
<text>[First paragraph text]</text>
</paragraph>
<paragraph>
<text>[Second paragraph text]</text>
</paragraph>
<paragraph>
<text>[Third paragraph text (if needed)]</text>
</paragraph>
</article>
</result>
```

7. After writing the article, double-check that you've included all the information from the event_info section.

Remember, your goal is to create a realistic and engaging news article based on the provided fictional event information while adhering to the newspaper's profile. Good luck!

H.29 Hallucination Removal Prompt in News Article Generation

You are tasked with improving a news article about a fictional event. Your goal is to ensure the article is faithful to the provided ground truth information while maintaining the general style of the original news article. Follow these instructions carefully:

1. First, review the ground truth information about the fictional named entities:

<entities>
<LOCATIONS>
{{LOCATIONS_XML}}
</LOCATIONS>

<PERSONS>
{{PERSONS_XML}}
</PERSONS>

<ORGANIZATIONS>
{{ORGANIZATIONS_XML}}
</ORGANIZATIONS>

<PRODUCTS>
{{PRODUCTS_XML}}
</PRODUCTS>

<ARTS>
{{ARTS_XML}}
</ARTS>

<EVENTS>
{{EVENTS_XML}}
</EVENTS>

<BUILDINGS>
{{BUILDINGS_XML}}
</BUILDINGS>

<MISCELLANEOUS>
{{MISCELLANEOUSS_XML}}
</MISCELLANEOUS>
</entities>

2. Next, review the ground truth information about the fictional event:

<event_info>
{{EVENT_INFO}}
</event_info>

3. Now, read the generated news article that needs to be revised:

<news-article>
{{CURRENT_NEWS_ARTICLE_XML}}
</news-article>

4. To revise the news article, follow these steps. Only revise the content of the article paragraphs. Do not revise the article title:

a) Analyze the style of the original news article. Pay attention to tone, vocabulary, and sentence structure. Any changes you make should maintain this style.

b) Carefully examine all factual statements in the article. Determine if each statement can be verified based on the ground truth information provided.

c) Identify any factual statements that are:
- Expressed as facts but are unverifiable based on the ground truth
- Incorrect according to the ground truth information

d) For each problematic statement, choose one of the following actions:
- Remove the unverifiable or incorrect factual statement entirely.
- Rephrase the statement to clearly indicate that it is not a verified fact (e.g., by using hedging language or attributing the information to an unnamed source).

e) Ensure that no information beyond what is provided in the ground truth is introduced as fact. However, you may include unverified information if it is presented as speculation, question, rumor or involves appropriate hedging.

f) Make any necessary adjustments to improve the flow and coherence of the article after your revisions.

g) Make minimal edits to the news article. Only make necessary revisions to avoid problematic factual statements. Explain and justify each change you make.

5. Output the revised news article in the following format:

```
<result>
<scratchpad>
(Plan your approach here, outlining the main changes you intend to make)
</scratchpad>
<headline>
(Write a revised headline that accurately reflects the content of the article)
</headline>
<article>
<paragraph>
<text>[First paragraph text]</text>
</paragraph>
<paragraph>
<text>[Second paragraph text]</text>
</paragraph>
<paragraph>
<text>[Third paragraph text (if needed)]</text>
</paragraph>
(Add more paragraphs as necessary, following the same format)
</article>
</result>
```

Remember to maintain the original style of the article while ensuring all factual statements are accurate according to the provided ground truth information.

H.30 Missing Information Adding Prompt in News Article Generation

You are an AI assistant tasked with improving a news article about a fictional event. Your goal is to ensure the article contains all specific details and information from a provided ground-truth outline of the fictional event. Follow these instructions carefully:

1. First, review the list of fictional named entities as background information:

```
<entities>
<LOCATIONS>
{{LOCATIONS_XML}}
</LOCATIONS>

<PERSONS>
{{PERSONS_XML}}
</PERSONS>

<ORGANIZATIONS>
{{ORGANIZATIONS_XML}}
</ORGANIZATIONS>
```

```
<PRODUCTS>
{{PRODUCTS_XML}}
</PRODUCTS>

<ARTS>
{{ARTS_XML}}
</ARTS>

<EVENTS>
{{EVENTS_XML}}
</EVENTS>

<BUILDINGS>
{{BUILDINGS_XML}}
</BUILDINGS>

<MISCELLANEOUS>
{{MISCELLANEOUSS_XML}}
</MISCELLANEOUS>
</entities>

2. Next, review the ground truth outline of the fictional event, including all details that must be communicated in the news article:

<event_info>
{{EVENT_INFO}}
</event_info>

3. Now, read the generated news article that needs to be revised:

<news-article>
{{CURRENT_NEWS_ARTICLE_XML}}
</news-article>

4. To revise the news article, follow these steps. Only revise the content of the article paragraphs. Do not revise the article title:
a) Analyze the style of the original news article. Pay attention to tone, vocabulary, and sentence structure. Any changes you make should maintain this style.
b) Go over each individual sentence from the ground truth outline.
c) Each sentence contains many details. Make sure that each of the details is communicated within the news article.
- The details do not need to be communicated verbatim. It is okay if the same content is communicated in different terms.
- Focus on all details from the sentence of the ground truth outline (numbers, dates, relations, relevant attributes and adjectives, etc). Consider every specific detail you can find.
d) Make subtle adjustments to the news article for each detail that is not yet communicated:
- Add the information with minimal edits
- Do not revise additional information from the news article such as speculations, rumors etc. Focus only on the missing information that must be integrated in the article.
e) Make any necessary adjustments to improve the flow and coherence of the article after your revisions.

5. Output the revised news article in the following format:

<result>
<scratchpad>
(Plan your approach here, outlining the main changes you intend to make)
</scratchpad>
<headline>
(Write a revised headline that accurately reflects the content of the article)
</headline>
<article>
<paragraph>
<text>[First paragraph text]</text>
</paragraph>
<paragraph>
<text>[Second paragraph text]</text>
</paragraph>
<paragraph>
<text>[Third paragraph text (if needed)]</text>
```

</paragraph>
(Add more paragraphs as necessary, following the same format)
</article>
</result>

Remember to maintain the original style of the article while ensuring all factual statements are accurate according to the provided ground truth information.

H.31 Named Entity Resolution Prompt in News Article Generation

You are an AI assistant tasked with processing news passages by identifying and marking named entities. Follow these instructions carefully:

First, review the list of fictional named entities provided below:

<entities>
<LOCATIONS>
{{USED_LOCATIONS_XML}}
</LOCATIONS>

<PERSONS>
{{USED_PERSONS_XML}}
</PERSONS>

<ORGANIZATIONS>
{{USED_ORGANIZATIONS_XML}}
</ORGANIZATIONS>

<PRODUCTS>
{{USED_PRODUCTS_XML}}
</PRODUCTS>

<ARTS>
{{USED_ARTS_XML}}
</ARTS>

<EVENTS>
{{USED_EVENTS_XML}}
</EVENTS>

<BUILDINGS>
{{USED_BUILDINGS_XML}}
</BUILDINGS>

<MISCELLANEOUS>
{{USED_MISCELLANEOUSS_XML}}
</MISCELLANEOUS>
</entities>

Next, you will process the following news article passages:

<passages>
{{PASSAGES_XML}}
</passages>

Your task is to process these passages by following these steps:

1. Carefully review the list of entities provided in the <entities> section. Each entity will have an associated ID.

2. Search the passages for all occurrences of each entity in the list.

3. For each occurrence found, replace it with the format: {phrase|ID}
Where "phrase" is exactly how the entity appears in the text (maintaining any abbreviations or variations), and "ID" is the entity's identifier from the entities list.

4. Maintain the original structure and formatting of the passages, only changing the entities as described.

5. After processing all entities, review the entire passage to ensure all occurrences have been properly marked and no entities were missed.

6. Output the processed passages, maintaining its original structure but with all entity occurrences replaced as instructed.

Important points to remember:
- Be thorough in your search for entities, including variations or partial mentions.
- Preserve the original text exactly as it appears, only adding the entity markup.
- Keep the COMPLETE ORIGINAL phrase that you are replacing with {phrase|ID}. The sentence should be identical to how it was before, except for the added markup.
- If an entity is referred to by full name, the "phrase" is the full name.
- If an entity is referred to by an abbreviation, the "phrase" is the used abbreviation.
- If an entity is referred to using parts of the full name, then the "phrase" would be the same parts of the full name.

Examples:
1. "Renowned novelist Elara Vance and celebrated philanthropist Rohan Kapoor exchanged vows."
Should be replaced with:
"Renowned novelist {Elara Vance|PERSON-1} and celebrated philanthropist {Rohan Kapoor|PERSON-2} exchanged vows."
(When Elara Vance has ID PERSON-1 and Rohan Kapoor has ID PERSON-2)

2. "Renowned novelist Elara and celebrated philanthropist R. Kapoor exchanged vows."
Should be replaced with:
"Renowned novelist {Elara|PERSON-1} and celebrated philanthropist {R. Kapoor|PERSON-2} exchanged vows."

3. "Anna Peters told Tim that he should stop talking."
should be written as:
"{Anna Peters|PERSON-3} told {Tim|PERSON-4} that {he|PERSON-4} should stop talking."
(When Anna Peters has ID PERSON-3 and Tim Laurens has ID PERSON-4 and is referred to here)

If multiple entities are referred to by the same word, use this format:
"{Both|PERSON-3,PERSON-4} liked the chocolate."

Format your output as follows:
- Enclose the entire processed news within <news> tags.
- Place each passage of the outline within separate <passage> tags.

Provide your final output without any additional commentary or explanations. Focus solely on processing the outline as instructed.

Provide all output in an overall <results> root node.

H.32 Answerability Filtering Prompt for Multi-hop Questions

You will receive evidence documents, a question, a date on which the question is asked, and answer options. Your task is to evaluate the evidence information, determine if it provides enough details to answer the question based on the date, and choose the correct answer.

Evidence:
{{EVENTS}}

Date of the question:
{{DATE}}

Question:
{{QUESTION}}

Answer Options:
{{ANSWERS}}

Instructions:

1. **Analyze the Evidence:**
- Carefully read all the provided evidence.
- Compare the information in the evidence with the question.
- Check if the combined evidence confirms all the necessary details to answer the question.

2. **Select an Answer:**
- Choose the correct answer if all necessary details are provided.
- If the evidence lacks information or any important detail is missing, select the option for "Unknown."

3. **Submit Your Answer:**
- Provide your response in JSON format as follows:
{
"justification": "Brief explanation (1-2 sentences)",
"answer_choice": "A single number corresponding to the chosen answer."
}

Output the JSON directly without any additional text or markdown syntax.

H.33 Answerability Filtering Prompt for Time-span Questions

You are tasked with evaluating evidence text to answer a question based on the provided information and a given date. Your goal is to determine if there's enough information to answer the question and select the correct answer option.

Here is the evidence you need to analyze:
{{EVENTS}}

The date on which the question is asked:
{{DATE}}

The question you need to answer:
{{QUESTION}}

The available answer options:
{{ANSWERS}}

Follow these steps to complete the task:

1. Carefully read and analyze all the provided evidence.
2. Compare the information in the evidence with the question.
3. Check if the combined information from the evidence confirms all the necessary details to answer the question.

4. Use a scratchpad to derive your answer:
<scratchpad>
- Identify all absolute start and end dates based on the evidence and any relevant assumptions from the question.
- Derive the answer based on these dates, if possible.
- Show your step-by-step reasoning process.
</scratchpad>

5. Select an answer:
- Choose the correct answer if all necessary details are provided.
- If the evidence lacks information or if any important detail is missing, select the option for "Unknown."
- If the question includes assumptions, treat them as facts.

6. Provide your response in JSON format as follows:
{
"scratchpad": "Your reasoning from the scratchpad",
"justification": "Brief explanation (1-2 sentences)",
"answer_choice": "A single number corresponding to the chosen answer."
}
Output the JSON directly without any additional text or markdown syntax.

H.34 Assumption Leakage Filtering Prompt for Time-span Questions

Given a question asked on a specific date, determine if there is sufficient evidence to answer the question.
If evidence is provided, base your answer on it. If no evidence is given, assess whether the question itself contains enough context to draw a conclusion.

Only provide an answer if the available evidence (or the question itself) is sufficient to determine it with certainty.
If the evidence is insufficient or unclear, select the answer option "Unknown" rather than making a guess.

```
Provide your response in JSON format as follows:
{ "scratchpad": "Your reasoning from the scratchpad",
"justification": "Brief explanation (1-2 sentences)",
"answer_choice": "A single number corresponding to the chosen answer."
}
Question Date: {{QUESTION_DATE}}
Question: {{QUESTION}}

Evidence:
{{EVIDENCE_TEXT}}

Answer Options:
{{ANSWER_OPTIONS}}
```

H.35 Multiple Choice Prompt for Experiments on NeoQA (Prompt 1)

Given the following news articles, the question, and the answer options, answer the question.
If the question cannot be answered with certainty based on the news articles, select the answer option "Unanswerable".

News Articles:
{{NEWS_ARTICLES}}

Question: {{QUESTION}}

Date of Question: {{DATE}}

Answer options:
{{ANSWERS}}

Select the answer option that correctly answers the question. If the question cannot be answered with certainty based on the news articles, choose "Unanswerable" (if it is one of the options). In the final line of your response, provide the number of the correct answer option using the format: "Answer: [answer number]" (for example, "Answer: X").

H.36 Multiple Choice Prompt for Experiments on NeoQA (Prompt 2)

You will receive news articles, a question, a date on which the question is asked, and answer options.
Your task is to evaluate the articles, determine if they provide enough information to answer the question based on the date, and choose the correct answer.

News Articles:
{{NEWS_ARTICLES}}

Question: {{QUESTION}}

Date of Question: {{DATE}}

Answer options:
{{ANSWERS}}

Instructions:
1. **Analyze the news articles:**
- Carefully read all the news articles.
- Compare the information in the articles with the question.
- Check if the combined information from the articles confirms all the details required to answer the question.

2. **Select an Answer:**
- Choose the correct answer if all necessary details are provided.
- If the articles lack information or any important detail is missing, select the option for "Unanswerable".

3. **Submit your Answer**
- Select the answer option that correctly answers the question. If the question cannot be answered with certainty based on the news articles, choose "Unanswerable" (if it is one of the options). In the final line of your response, provide the number of the correct answer option using the format: "Answer: [answer number]" (for example, "Answer: X").

H.37 Multiple Choice Prompt for Experiments on NeoQA (Prompt 3)

You will receive news articles, a question, a date on which the question is asked, and answer options.
Your task is to evaluate the articles, determine if they provide enough information to answer the question based on the date, and choose the correct answer.

News Articles:
{{NEWS_ARTICLES}}

Question: {{QUESTION}}

Date of Question: {{DATE}}

Answer options:
{{ANSWERS}}

Instructions:
1. **Analyze the news articles:**
- Carefully read all the news articles.
- Compare the information in the articles with the question.
- Check if the combined information from the articles confirms all the details required to answer the question.
- Ensure that the question does not contain contradictory information compared to the provided news articles. Select the "Unanswerable" option if it does.
- Verify that the information in the news articles is sufficient to answer the question with certainty. If you cannot answer the question with certainty based on the evidence, select the "Unanswerable" option.
- The news articles may not be in the correct temporal order.
- If the question mentions an "event date", this refers to the date of the news article.
- Unless otherwise stated, you can assume that each news article reports events that occurred on the date of the article.

2. **Select an Answer:**
- Choose the correct answer if all necessary details are provided.
- If the articles lack information or any important detail is missing, select the option for "Unanswerable".

3. **Submit your Answer:**
- Select the answer option that correctly answers the question. If the question cannot be answered with certainty based on the news articles, choose "Unanswerable" (if it is one of the options). In the final line of your response, provide the number of the correct answer option using the format: "Answer: [answer number]" (for example, "Answer: X").

H.38 Multiple Choice Prompt for Experiments on NeoQA (Prompt 4)

You will receive news articles, a question, a date on which the question is asked, and answer options.
Your task is to evaluate the articles, determine if they provide enough information to answer the question based on the date, and choose the correct answer.

News Articles:
{{NEWS_ARTICLES}}

Question: {{QUESTION}}

Date of Question: {{DATE}}

Answer options:
{{ANSWERS}}

Instructions:
1. **Analyze the news articles:**
- Carefully read all the news articles.
- Compare the information in the articles with the question.
- Check if the combined information from the articles confirms all the details required to answer the question.
- Ensure that the question does not contain contradictory information compared to the provided news articles. Select the "Unanswerable" option if it does.
- Verify that the information in the news articles is sufficient to answer the question with certainty. If you cannot answer the question with certainty based on the evidence, select the "Unanswerable" option.
- The news articles may not be in the correct temporal order.
- If the question mentions an "event date", this refers to the date of the news article.
- Unless otherwise stated, you can assume that each news article reports events that occurred on the date of the article.

2. **Double-check the details:**
- Use only the information provided in the news articles.
- Avoid assumptions beyond what is explicitly stated.
- Do not make guesses. Only provide an answer if the information in the article is enough to answer the question with certainty. If it's not, select the "Unanswerable" option.

3. **Select an Answer:**
- Choose the correct answer if all necessary details are provided.
- If the articles lack information or any important detail is missing, select the option for "Unanswerable".

4. **Submit your Answer:**
- Select the answer option that correctly answers the question. If the question cannot be answered with certainty based on the news articles, choose "Unanswerable" (if it is one of the options). In the final line of your response, provide the number of the correct answer option using the format: "Answer: [answer number]" (for example, "Answer: X").

H.39 Multiple Choice Prompt for Experiments on NeoQA (Prompt 5)

```
You will receive news articles, a question, a date on which the question is asked, and answer options.
Your task is to evaluate the articles, determine if they provide enough information to answer the question based on the date, and choose the correct answer.

News Articles:
{{NEWS_ARTICLES}}

Question: {{QUESTION}}

Date of Question: {{DATE}}

Answer options:
{{ANSWERS}}

**Instructions:**
1. **Analyze the news articles:**
- Carefully read all the news articles.
- Compare the information in the articles with the question.
- Check if the combined information from the articles confirms all the details required to answer the question.
- Ensure that the question does not contain contradictory information compared to the provided news articles. Select the "Unanswerable" option if it does.
- Verify that the information in the news articles is sufficient to answer the question with certainty. If you cannot answer the question with certainty based on the evidence, select the "Unanswerable" option.
- The news articles may not be in the correct temporal order.
- If the question mentions an "event date", this refers to the date of the news article.
- Unless otherwise stated, you can assume that each news article reports events that occurred on the date of the article.

2. **Double-check the details:**
- Use only the information provided in the news articles.
- Avoid assumptions beyond what is explicitly stated.
- Do not make guesses. Only provide an answer if the information in the article is enough to answer the question with certainty. If it's not, select the "Unanswerable" option.
- Make sure that all the necessary information from the question is present in the news article. For each detail in the question, write down how you verified it against the articles, along with your conclusion. If any important details are missing and it's unclear whether the article fully supports the question, select the "Unanswerable" option. - Use only the information provided in the news articles.

3. **Select an Answer:**
- Choose the correct answer if all necessary details are provided.
- If the articles lack information or any important detail is missing, select the option for "Unanswerable".

4. **Submit your Answer:**
- Select the answer option that correctly answers the question. If the question cannot be answered with certainty based on the news articles, choose "Unanswerable" (if it is one of the options). In the final line of your response, provide the number of the correct answer option using the format: "Answer: [answer number]" (for example, "Answer: X").
```