D Analyzing the OpenForesight Dataset

Table 3: Breakdown of source news articles by news outlet.

Source	Articles (%Total)
Forbes	110,103 (44.3%)
The Hindustan Times	80,000 (32.2%)
The Irish Times	29,546 (11.9%)
Deutsche Welle (DW)	21,317 (8.6%)
Cable News Net (CNN)	7,355 (3.0%)
Total	**248,321 (100%)**

News Corpus Details. We drew from a heterogeneous pool of news articles diversified by geography, time, and topic to ensure broad coverage. We collected a English-language articles from outlets including Forbes, Hindustan Times, The Irish Times, Deutsche Welle, and CNN. They were selected by first collecting a corpus of approximately 250,000 articles spanning from June 2023 to April 2025, encompassing major large-scale events across sports, geopolitics, local news, crime, entertainment, and the arts. Then, performing de-duplication and filtering for language, text availability, and valid dates, we retained approximately 248,000 articles for the question generation phase. Table 3 details the distribution across news outlets for our corpus.

The overall dataset creation process costed us 3000$ with training set costing 2200$ (using DeepSeek-v3) while creating the test set costed 800$ (using o4-mini-high and grok-4.1-fast with search tool).

D.1 Analysis of Best Question Selection

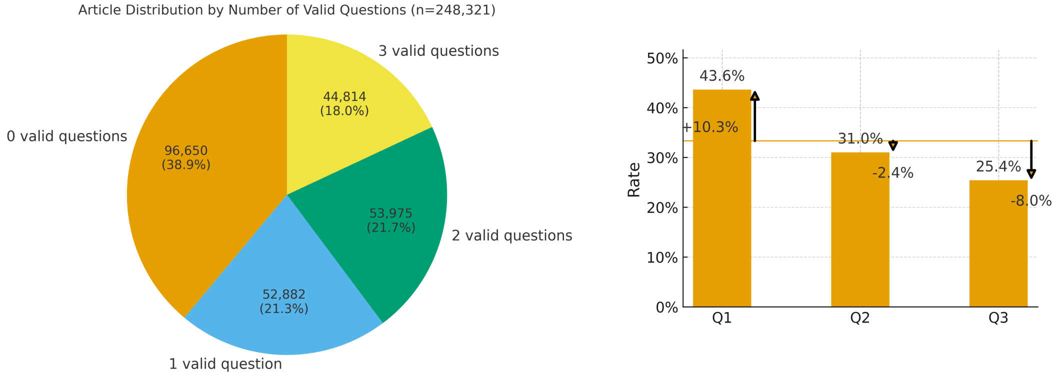


Figure 17: **Data Distribution of Questions in** OpenForesight. (Left) Distribution of number of questions selected after filtering from articles (Right) We show the number of questions generated, and the proportion of the first, second and third generated question being picked as the final "best question".

As Figure 17 (left) illustrates, the generation process yielded mixed results. Post-processing, 39% of source articles failed to produce any valid questions. Among the surviving articles, 21% yielded exactly one valid question, which we retained. For the 61% of articles producing multiple valid questions, we employed another LLM to identify the best candidate based on global relevance, specificity, and unambiguity. Our analysis of selected questions, shown in Figure 17 (right) reveals a selection bias:

- Question 1: Selected 43.6% of the time (10.3% above random).
- Question 2: Selected 31% of the time (2.4% below random).
- Question 3: Selected 26.7% of the time (6.7% below random).

This trend suggests that initial generation attempts (question positioned earlier) frequently produce higher-quality results. Figure 5 provides qualitative examples of these generated questions.

	Name(s)	Location	Country	Title	Team name	Color	Organization	Currency	Brand name	Month
Count	32,213	14,337	2,579	2,479	1,445	1,047	1,030	877	779	730
Share	44.8%	20.0%	3.6%	3.5%	2.0%	1.5%	1.4%	1.2%	1.1%	1.0%

Table 4: Top ten answer types of the questions in our curated dataset. These ten categories cover 80.1% of our training dataset.

D.2 Distribution of Answer Types

Table 4 categorizes the answer types within the training data. Two categories dominate the dataset:

- People and Places (65%): Names of individuals constitute nearly 45% of the answers, while locations account for 20%.
- Miscellaneous Entities (35%): The remainder consists of teams, countries, organizations, colors, and similar entities.

Question	**Background**	**Resolution (trigger & deadline)**	**Answer Type**	**Answer**	**Source**
Host country of COP30 (Nov 2025)?	UNFCCC COP venue rotates among regions.	Host confirmed by UNFCCC/organizers; no later than COP30 start (Nov 2025).	string (country)	Brazil	DW: link
Release month of Marvel's *Fantastic Four* (2025)?	Reboot announced with lead cast; 2025 release slated.	Month confirmed by Marvel/Disney; by Dec 2025.	string (month)	July	Forbes: link
First state to require Ten Commandments in public classrooms (by 2025)?	Several U.S. states advance religion-in-school measures.	First state enacts requirement; by Dec 31, 2025.	string (state name)	Louisiana	Forbes: link
African host of G20 Summit (Nov 2025)?	G20 presidency rotates; South Africa presiding from Dec 2024.	G20/host government confirms location; by Nov 2025.	string (country)	South Africa	DW: link
Recipient of Lesotho–Botswana Transfer Scheme (by 2025)?	Regional pipeline to pump water from Lesotho via SA.	ORASECOM or governments confirm recipient; by 2025.	string (country name)	Botswana	DW: link

Table 5: Five succinct forecasting questions spanning climate, entertainment, law, geopolitics, and infrastructure; selected for brevity and diverse sources (DW, Forbes). Each row lists the question (summarized here for conciseness), short background, resolution trigger with deadline, answer type, ground-truth answer, and citation.

D.3 Test Set

For the test set, we first prepare an initial set of 1000 questions generated using `o4-mini-high` model as described in Section 6. We next performed additional filtering steps to retain high-quality future-facing questions:

1. We removed any potentially unanswerable questions (noise) by keeping only those which `grok-4.1-fast` could successfully answer majority of the times (run repeatedly, $n = 5$) with search tool access. This filtered out 15% of the questions.
2. To address the issue of late reporting in news outlets, we again use `grok-4.1-fast` with search tool to find the **earliest resolution date** for a given question. This is important to prevent leakage from retrieving articles with the true answer. We retain only those questions with resolution date after May 2025. In Figure 18b, we report the number of questions per news source for whom the generated resolution date was within 1 month period of the date found by `grok-4.1-fast`. We notice an average of 70% questions have resolution date even within the 1 month period.

Finally, we manually filter the remaining questions to meet our quality checks like:

1. Question may have multiple possible correct answers.
2. Question actually resolves in the future (after September 2025) for which the article reports the scheduled/planned place/event/etc.
3. Question being irrelevant because it is too niche to a certain place or locality.
4. Question is about something which is already established (known).

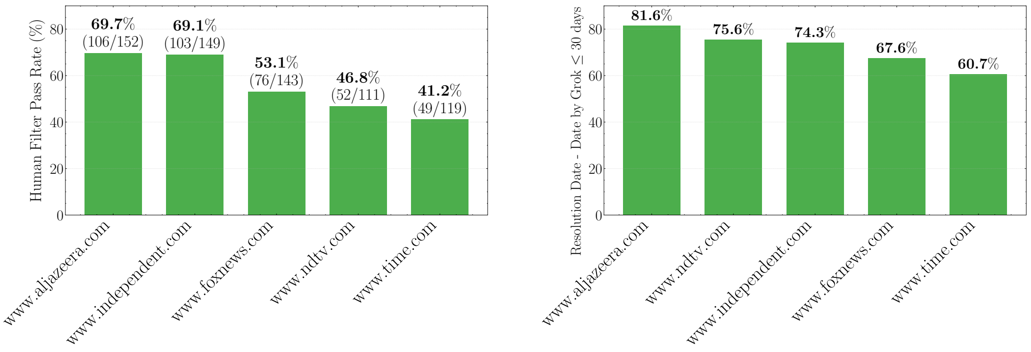
We provide the full guidelines we followed for manual filtering in the box below. This resulted in a final test set of 302 questions which we use for reporting the results.

Manual Filtering Guidelines

```
**Task:** You will be given (i) a news article and (ii) a question that is
    supposed to be answerable from that article. Your job is to decide whether
    to KEEP the question in the test set or REJECT it during manual filtering.

**How to review:**
1. Read the full question (Title, Background, Resolution Criteria, and the
    proposed Answer).
2. Apply the rejection criteria below. If **any** criterion triggers, REJECT the
    question. Refer to the article text if required.

------------------------------------------------------------
```



(a) Manual filter pass rate per news source.

(b) Proportion of questions whose resolution date is within 1 month of the resolution date found by grok-4.1-fast.

Figure 18: Filtering pass rate of forecasting questions across news sources.