

**\*\*REJECT the question if it fails ANY of the following criteria:\*\***

- 1) **\*\*Multiple possible correct answers (Ambiguous / not uniquely resolvable)\*\***
  - Reject if the question could reasonably have more than one correct answer, even after reading the article.
  - Examples:
    - "Which company invested in ABC in July 2025?" when multiple companies are mentioned investing in ABC in July 2025.
    - "Who announced a new partnership?" when the article lists several partnerships.
- 2) **\*\*Resolves in the future (after September 2025)\*\***
  - Reject if the question's true resolution depends on an event that happens **\*\*after 2025-09-30\*\***, and the article only reports a **\*plan/schedule/announcement\*** rather than the event actually occurring.
  - Examples:
    - Article: "The ABC event is scheduled for Jan 2026 in PQR venue."  
Question: "Where will event ABC be held in Jan 2026?" -> Reject (the question resolves only after the event actually takes place as the venue can change last minute).
- 3) **\*\*Too niche / overly local / low general relevance\*\***
  - Reject if the question is narrowly tied to a very specific locality, institution, or small community such that it is not broadly meaningful for a general test set.
  - Examples:
    - A minor municipal policy affecting only one small village.
    - A local event with no wider regional/national/global significance.
- 4) **\*\*Already established (Known / not a meaningful forecast target)\*\***
  - Reject if the question is about something that is already settled or widely known at the time (i.e., not an uncertain outcome).
  - Examples:
    - "Who stepped down from the role of CEO of Company ABC in August 2025?" (requires knowing the current CEO which generally doesn't change frequently and is often stable over multiple years)
    - Long-established facts, definitions, or historical constants.

---

**\*\*Decision format (for annotation):\*\***

- If any criterion triggers, then **\*\*REJECT\*\***.
- Otherwise, **\*\*KEEP\*\***.

**\*\*Question:\*\***

{questions\_text}

<toggle> **\*\*Source Article:\*\*** {source\_article} </toggle>

## E Qualitative Analysis

### E.1 Qualitative Analysis of Final Answers

We manually annotated responses to 207 questions by both the initial Qwen3-8B thinking model and the trained OpenForecaster8B on the Guardian validation set. Using this set, we found that the agreement between the two models used for grading, Llama 4 Scout and Qwen3 4B is  $\sim 97\%$ , and we agree with their grading in over 95% of cases. This confirms the reliability of automatic answer matching based evaluation.

In Table 6, we analyze the domains (by news section) in which our trained model improves. We find significant improvements in the World, Australian, and US news sections, with no significant change for sports. This suggests our model may not yet perform well on sports-heavy prediction markets like Kalshi.

Domain	<i>n</i>	Before	After	$\Delta$
world	20	21.7	33.3	+11.6
australia-news	15	35.6	42.2	+6.7
us-news	21	41.3	44.4	+3.2
sport	37	43.2	43.2	+0.0
football	30	34.4	33.3	-1.1

Table 6: Avg@3 by domain ( $n \geq 10$ ).

In Table 7, we analyze change in performance by question type, finding significant improvements on questions of the form “what”, “which”, and “who”, while a slight regression in performance on location questions (“where”).

Question form	<i>n</i>	Before	After	$\Delta$
what	25	14.7	29.3	+14.7
which	98	45.2	51.4	+6.1
who	60	27.8	33.9	+6.1
other	10	40.0	43.3	+3.3
where	14	47.6	45.2	-2.4

Table 7: Avg@3 by question form ( $n \geq 10$ ).

Below, we present qualitative examples where our training improves and worsens predictions compared to the original model.

#### *Qualitative examples (improved; first sample)*

- **Q:** Who will be wearing the yellow jersey in the general classification at the end of stage eight of the 2025 Tour de France?  
**Truth:** Tadej Pogačar  
*Before:* Jonas Vingegaard (p=0.10)  
*After:* Tadej Pogačar (p=0.60)
- **Q:** Who will withhold a resolution from the U.S. House floor to force a vote on releasing the Epstein documents by July 25, 2025?  
**Truth:** Mike Johnson  
*Before:* Pam Bondi (p=0.30)  
*After:* Mike Johnson (p=0.60)

#### *Qualitative examples (regressed; first sample)*

- **Q:** Who will be the Democratic Party’s nominee for New York City mayor in the November 2025 general election?  
**Truth:** Zohran Mamdani  
*Before:* Zohran Mamdani (p=0.60)  
*After:* Andrew Cuomo (p=0.40)

- Q: Who will post the lowest first-round score among Rory McIlroy, Scottie Scheffler and Viktor Hovland at the 2025 Scottish Open?

**Truth:** Viktor Hovland

*Before:* Viktor Hovland (p=0.60)

*After:* Scottie Scheffler (p=0.40)

## E.2 Reasoning Evolution During Training

We now provide a few qualitative examples of how the model reasoning evolves when training on OpenForesight. We ask the model itself to provide a summary of its reasoning (after the `<think></think>` tags), and show this summary below for the initial Qwen3-8B model, followed by training checkpoints.

### E.2.1 Example 1: Model stays incorrect but learns to hedge

**Question:** Which country will face a new 25% US tariff in a trade announcement by 30 July 2025?

**Background:** The US administration is evaluating other countries' trade barriers and is expected to announce new tariffs by the end of July 2025.

**Resolution Criteria:**

**Source of Truth:** Official tariff announcement issued by the White House or the Office of the United States Trade Representative.

**Resolution Date:** 30 July 2025, the date by which the 25% tariff announcement is expected.

**Accepted Answer Format:** The full name of the country as stated in the official announcement.

**Resolution answer:** India

*Checkpoint 0: score=-0.722, status=wrong*

`<answer>South Korea</answer> <probability>0.85</probability>` \*\*Reasoning\*\*: - \*\*Article 2\*\* (May 25, 2025) explicitly mentions a \*\*25% tariff\*\* on \*\*Samsung\*\* (South Korea) and other smartphone manufacturers. This directly ties the 25% tariff to \*\*South Korea\*\*. - While the EU is mentioned with a \*\*50% tariff\*\* in the same article, the \*\*25% tariff\*\* specifically targets South Korea's exports. - Other articles reference tariffs on China, the EU, or automakers, but none specify a \*\*25% tariff\*\* on a country beyond South Korea. - The resolution date (July 30, 2025) aligns with the timeline of the May 25 announcement, as no newer data is provided. \*\*Confidence\*\*: 85% (strong evidence from Article 2, though no post-May 25 updates confirm further announcements).

*Checkpoint 400: score=-0.360, status=wrong*

The retrieved articles indicate ongoing US tariff announcements by the Trump administration in 2025. Article 2 explicitly mentions a 25% tariff on South Korea for smartphone manufacturers like Samsung, while Article 4 references 25% tariffs on steel and aluminum imports from Canada and Mexico. Article 3 also notes a 25% tariff on foreign automakers, which could target countries like South Korea or others, but South Korea is specifically named in Article 2 for a 25% rate. Given the question's focus on a "new" 25% tariff announcement by July 30, 2025, and the patterns in the articles, South Korea appears as a likely candidate, as it has already faced such announcements and fits the criteria. However, the question implies a potential new announcement, and other countries like Canada or Mexico could also be possibilities based on Article 4, introducing uncertainty. My confidence is moderate due to the reliance on past announcements and the possibility of new developments by the resolution date. `<answer>South Korea</answer> <probability>0.6</probability>`

*Checkpoint 800: score=-0.090, status=wrong*

The retrieved articles indicate a pattern of US tariffs targeting major trade partners like China, India, the EU, Canada, Mexico, and South Korea. Article 4 specifically mentions a 25% tariff on steel and aluminum imports, which historically targets countries such as China and those in the EU. Article 2 highlights a 25% tariff on South Korean products