E.3 Systematic Failure Modes in Model Reasoning

`OpenForecaster` 8B achieved an accuracy of 33% on our test set consisting of 302 questions. Through manual analysis of the failure cases, we identified three systematic failure modes in the model's reasoning process.

1. Missing Information in Retrieved Articles The most common failure mode occurs when retrieved articles lack the specific information needed. The model acknowledges this but falls back to general knowledge or educated guesses rather than expressing high uncertainty.

Example 1: "By what date will the New York City Board of Elections certify the official results of the Democratic mayoral primary?"

- **Correct answer:** July 15
- **Model answer:** August 15 (probability: 0.2)
- **Model's reasoning:** "The provided news articles discuss the New York City Democratic mayoral primary, including polling data, candidate strategies, and the election date of June 24, 2025. However, they do not specify the certification process or timeline for the New York City Board of Elections. Based on general knowledge of past NYC elections, certification typically occurs 2-6 weeks after the primary to allow for vote counting, ballot challenges, and legal reviews. For instance, in the 2013 primary, certification was completed by July 2. Given the late summer 2025 resolution criteria, a reasonable estimate is mid-August, but without specific historical data from the articles, this is uncertain. I am guessing August 15, 2025, as it aligns with typical procedures, but my confidence is low due to the lack of direct evidence."
- **Analysis:** The model correctly identifies that articles lack the certification date but incorrectly infers from general knowledge. It acknowledges uncertainty (probability 0.2) yet provides a specific date. The inference from "2-6 weeks after June 24" to "mid-August" is flawed, as July 15 falls within this range.

2. Over-reliance on General Knowledge When articles lack specific information, the model makes educated guesses based on general knowledge, historical patterns, or analogies rather than acknowledging the answer cannot be determined from the provided context.

Example 1: "What will be the official name of Donna D'Errico's interactive AI voice service by mid-July 2025?"

- **Correct answer:** Call Donna D.
- **Model answer:** Donna Voice (probability: 0.1)
- **Model's reasoning:** "The provided articles do not mention Donna D'Errico or her specific AI voice service, making it impossible to draw direct insights. However, patterns from the articles, such as the naming of AI voice services (e.g., Melania Trump's "Melania: The Audiobook"), suggest that such services often incorporate the individual's name followed by descriptors like "Voice" or "AI." Based on this analogy, I am guessing the name might be "Donna Voice" or something similar. This is a speculative prediction, as there's no concrete information, leading to low confidence."
- **Analysis:** The model creates a naming pattern from one unrelated example (Melania Trump's service) and applies it incorrectly. It recognizes the speculation (probability 0.1) but the actual name "Call Donna D." follows a different pattern entirely.

Example 2: "What name will the new Pope take upon his election on May 8, 2025?"

- **Correct answer:** Leo XIV
- **Model answer:** Pius (probability: 0.1)

- **Model's reasoning:** "The provided articles do not specify the name of the next Pope or any potential candidates' chosen names. They focus on Pope Francis's health, the conclave process, and recent appointments of cardinals. Historically, new Popes select names that are traditional, often drawing from saints or biblical figures, but without specific information from the articles or external knowledge, any prediction is highly uncertain. Based on patterns from past Popes (e.g., Francis, Benedict), I am guessing "Pius" as a common regnal name, but this is purely speculative and not supported by the articles. Therefore, my confidence in this answer is very low."
- **Analysis:** The model relies entirely on general knowledge about papal naming conventions rather than specific information. It correctly identifies high uncertainty (probability 0.1) though. The reasoning conflates "common regnal names" with what the next Pope will actually choose.

3. Entity Confusion: Selecting Wrong Person/Place/Organization When multiple entities are mentioned, the model frequently selects the most prominent or frequently mentioned one rather than the correct answer, even when both are present in the context.

Example 1: "Who will Chelsea face in the round of 16 of the 2025 FIFA Club World Cup?"

- **Correct answer:** Benfica
- **Model answer:** Bayern Munich (probability: 0.4)
- **Model's reasoning:** "The provided articles detail the groups for the 2025 FIFA Club World Cup, with Group C consisting of Bayern Munich, Benfica, Boca Juniors, and Auckland City, and Group D including Chelsea, Flamengo, Esperance Sportive de Tunisie, and a replacement (possibly Los Angeles FC or Club America based on Article 1). The question specifies that the round of 16 matchup is between the runner-up of Group D and the winner of Group C. Without the actual group stage results, I must make an educated guess based on team strengths. Bayern Munich is a strong candidate for the winner of Group C, as they are the defending champions and one of the top teams. Chelsea, in Group D, could plausibly be the runner-up, given their history, but this is uncertain. The replacement team adds further uncertainty, as it could affect group dynamics."
- **Analysis:** The model selects the more prominent team (Bayern Munich) from Group C rather than correctly identifying which team would actually win the group. It uses team prominence ("defending champions," "top teams") as a proxy for group stage results, ignoring that the question requires specific match outcomes that aren't in the articles.

F Prompts

F.1 Prompt Templates

F.1.1 Question Creation Pipeline

Stage 1 — Question Generation (Requires: self.num_questions_per_article > 1)

```
**Task:** Based on the provided news article, generate
    {self.num_questions_per_article} high-quality, DIVERSE forecasting questions
    which have a short answer (1 - 3 words), using the XML format specified
    below.
Each forecasting question should be posed in a way to predict future events.
    Here, the predictor will have a knowledge cutoff before the article is
    published and no access to the article, so a forecasting question has to be
    posed about information explicitly stated in the article. The question
    should be stated in a forward-looking manner (towards the future).
The correct answer should be a specific, short text response. The answer should
    be a WELL DEFINED, SPECIFIC term which the answerer can come up with on its
    own, without access to the news article.

**Example Format**:
<q1>
<question_id>0</question_id>
<question_title>Who will win the Nobel Prize in Literature in
    2016?</question_title>
<background>Question Start Date: 10th January 2016. The Nobel Prize in
    Literature is awarded annually by the Swedish Academy to authors for their
    outstanding contributions to literature.</background>
<resolution_criteria>
<ul>
    <li>
      <b>Source of Truth</b>: The question will resolve when the Swedish Academy
    publicly announces the official 2016 Nobel Prize in Literature
    laureate(s)---typically via a press release on NobelPrize.org (expected on
    or about October 13, 2016).
    </li>
    <li>
      <b>Resolution Date</b>: The resolution occurs on the calendar date when
    the 2016 laureate(s) are formally named
      (typically mid-October 2016).
    </li>
    <li>
      <b>Accepted Answer Format</b>: The full name of the laureate exactly as
    given in the announcement should be provided. If more than one person shares
    the prize, all names must be listed in the same order as the official
    communiqu{\'e}.
    </li>
</ul>
</resolution_criteria>
<answer>Bob Dylan</answer>
<answer_type>String (Name)</answer_type>
</q1>

The question should follow the structured guidelines below.

### **Guidelines for Creating Short Answer Forecasting Questions**

**Title Question Guidelines**
- **Quality**: The question should be of HIGH QUALITY and hard to answer without
    access to the article. It should not be about any minute details in the
    article. THE QUESTION SHOULD BE SUCH THAT ITS ANSWER REVEALS A KEY PIECE OF
    INFORMATION, FROM THE ARTICLE, WHICH HAS MAXIMAL IMPACT.
```