```
You are an expert forecaster evaluating prediction market questions.
Given a question and its description, predict whether the outcome
will be "yes" or "no".

You must respond in this exact format:
<think>
[Your reasoning about the prediction, considering base rates,
relevant factors, and uncertainty]
</think>
<answer>[yes or no]</answer>
<confidence>[a number from 0 to 100 representing your confidence
that the answer is "yes"]</confidence>

Be calibrated: if you're 70% confident, you should be correct
about 70% of the time on similar questions.

USER PROMPT:
Question: {question}

Description: {description}
```

The explicit calibration instruction ("Be calibrated: if you're 70% confident, you should be correct about 70% of the time") makes the observed miscalibration particularly notable—models fail to achieve calibration even when directly instructed to do so.

C Dataset Creation Details

The KalshiBench dataset was created through the following pipeline:

1. **Raw data collection**: Query Kalshi API for all resolved binary contracts.
2. **Temporal filtering**: Retain only contracts resolving after October 1, 2025.
3. **Deduplication**: Limit to 2 questions per series_ticker to reduce redundancy while preserving within-series diversity.
4. **Quality filtering**: Remove contracts with ambiguous resolution criteria or missing ground truth.
5. **Schema standardization**: Map to unified schema with fields: id, question, description, category, close_time, ground_truth.

The final dataset contains 300 questions across 13 categories, with category entropy of 3.01 bits (maximum possible: 3.70 bits), indicating reasonable diversity.