3.4 Deduplication and Quality Control

Raw prediction market data contains redundant questions (e.g., daily instances of recurring markets). We limit to 2 questions per series ticker to preserve diversity while reducing redundancy. All questions include detailed resolution criteria in the description field, ensuring unambiguous ground truth.

4 Methodology

4.1 Models Evaluated

We evaluate five frontier models representing diverse architectures and training approaches:

Table 3: Models evaluated in KalshiBench. All models have knowledge cutoffs at or before October 2025.

Model	Provider	Knowledge Cutoff	Notes
Claude Opus 4.5	Anthropic	April 2025	Flagship model
GPT-5.2-XHigh	OpenAI	October 2025	Extended reasoning
DeepSeek-V3.2	DeepSeek	October 2025	Open-weight
Qwen3-235B-Thinking	Alibaba	June 2025	Reasoning-enhanced
Kimi-K2	Moonshot	June 2025	Reasoning-enhanced

4.2 Evaluation Protocol

Each model receives a structured prompt containing the prediction market question and resolution criteria. The system prompt explicitly instructs models to be calibrated:

```
System: 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.
```

The user message then provides the specific question and description. Notably, the prompt explicitly instructs models to "be calibrated," making the observed miscalibration a failure to follow instructions rather than mere absence of guidance.

We use temperature 0.7 for standard models and temperature 1.0 with extended reasoning for GPT-5.2-XHigh, following provider recommendations.

4.3 Metrics

Classification Metrics. We report accuracy, precision, recall, and macro-F1 for binary classification performance.

Brier Score. The Brier score [Brier, 1950] measures the mean squared error of probability predictions:

$$\text{BS} = \frac{1}{N}\sum_{i=1}^{N}(p_i - y_i)^2 \tag{2}$$

where p_i is the predicted probability and $y_i \in \{0, 1\}$ is the outcome. Lower is better (0 = perfect, 1 = worst possible).

Intuition: The Brier score can be interpreted as follows:

- **0.00**: Perfect predictions—100% confidence on all correct answers
- **0.25**: Random guessing (50% confidence on everything)—the expected score of a completely uninformed predictor on balanced binary outcomes
- **0.20**: Good calibration—roughly equivalent to human forecasters on prediction markets
- **0.33**: Poor calibration—equivalent to always predicting 42% (the base rate) with uniform 75% confidence
- **1.00**: Maximally wrong—100% confidence on all incorrect answers

For context, human superforecasters typically achieve Brier scores of 0.15–0.20 [Tetlock & Gardner, 2015], while the aggregate "wisdom of crowds" on prediction markets often achieves 0.12–0.18.

Brier Skill Score. The Brier Skill Score (BSS) measures improvement over a baseline that always predicts the base rate:

$$\text{BSS} = 1 - \frac{\text{BS}}{\text{BS}_{\text{climatology}}} \tag{3}$$

where $\text{BS}_{\text{climatology}} = \bar{y}(1 - \bar{y})$ for base rate $\bar{y}$. Positive values indicate improvement over the base rate.

Expected Calibration Error (ECE). ECE [Naeini et al., 2015] measures the average gap between confidence and accuracy:

$$\text{ECE} = \sum_{b=1}^{B}\frac{|B_b|}{N}\left|\text{acc}(B_b) - \text{conf}(B_b)\right| \tag{4}$$

where predictions are binned by confidence into B bins.

Maximum Calibration Error (MCE). MCE captures the worst-case calibration in any single bin:

$$\text{MCE} = \max_{b\in\{1,\ldots,B\}}\left|\text{acc}(B_b) - \text{conf}(B_b)\right| \tag{5}$$

Overconfidence Rate. We define overconfidence rate at threshold τ as the fraction of incorrect predictions among those with confidence $> \tau$:

$$\text{OCR@}\tau = \frac{|\{i : p_i > \tau \wedge \hat{y}_i \neq y_i\}|}{|\{i : p_i > \tau\}|} \tag{6}$$

5 Results

5.1 Main Results

Table 4 presents comprehensive results across all models and metrics.

Key Finding 1: Systematic Overconfidence. All models exhibit substantial calibration errors, with ECE ranging from 0.120 to 0.395. Even the best-calibrated model (Claude Opus 4.5) shows a 12-percentage-point average gap between confidence and accuracy.

Table 4: Main results on KalshiBench (300 questions). Best values in **bold**. Claude Opus 4.5 achieves best performance on both accuracy and calibration metrics. Notably, the reasoning-enhanced GPT-5.2-XHigh shows the worst calibration despite comparable accuracy.

	Classification			Calibration			
Model	Acc	F1	$F1_{yes}$	Brier ↓	BSS ↑	ECE ↓	MCE ↓
Claude Opus 4.5	**69.3**	**0.676**	**0.600**	**0.227**	**0.057**	**0.120**	**0.246**
Kimi-K2	67.1	0.633	0.515	0.347	-0.446	0.298	0.570
Qwen3-235B	65.7	0.607	0.466	0.346	-0.437	0.297	0.479
GPT-5.2-XHigh	65.3	0.599	0.453	0.433	-0.799	0.395	0.622
DeepSeek-V3.2	64.3	0.614	0.507	0.339	-0.407	0.284	0.630

Table 5: Confidence analysis across models. All models show higher confidence when wrong than would be appropriate for well-calibrated predictions. Overconfidence rates at high confidence levels (80%+, 90%+) are alarmingly high.

Model	**Avg Conf**	$\mathbf{Conf_{wrong}}$	**OCR@70**	**OCR@80**	**OCR@90**
Claude Opus 4.5	73.8%	71.0%	27.1%	23.1%	**20.8%**
DeepSeek-V3.2	73.7%	69.2%	**24.7%**	**23.6%**	**14.7%**
Kimi-K2	79.4%	76.3%	25.9%	29.9%	31.1%
GPT-5.2-XHigh	80.1%	76.9%	30.3%	28.3%	27.7%
Qwen3-235B	81.7%	80.4%	32.3%	32.6%	32.4%

Key Finding 2: Most Models Fail to Beat the Base Rate. Only Claude Opus 4.5 achieves a positive Brier Skill Score (0.057), indicating it marginally outperforms simply predicting the 40% base rate. All other models have negative BSS, meaning their probability estimates are *worse than uninformed guessing*.

Key Finding 3: Reasoning Enhancements Hurt Calibration. Counterintuitively, GPT-5.2-XHigh (with extended reasoning) shows the worst calibration (ECE=0.395, BSS=-0.799) despite using 26× more output tokens (∼2M vs ∼138K for Claude). Enhanced reasoning appears to increase confidence without proportional accuracy gains.

5.2 Confidence Analysis

Table 5 reveals troubling patterns in model confidence:

- **High baseline confidence**: Models average 74-82% confidence, far exceeding the 65-69% accuracy range.
- **Confidence when wrong**: Models maintain 69-80% confidence even on incorrect predictions, indicating poor uncertainty awareness.
- **Extreme overconfidence**: At the 90%+ confidence level, models are wrong 15-32% of the time. A well-calibrated model should be wrong <10%.

5.3 Reliability Diagrams

A reliability diagram plots predicted confidence against actual accuracy across binned predictions. A perfectly calibrated model follows the diagonal: when it expresses 70% confidence, it should be correct 70% of the time. Table 6 presents complete reliability data for all five models across all 10 confidence bins.

Several patterns emerge from the reliability analysis:

Claude Opus 4.5 shows the best calibration overall, with relatively small gaps in most bins. However, even Claude becomes overconfident at high confidence levels: at 90%+ confidence (20 predictions), accuracy is only 70%, yielding a +24.6% gap.