# Dependence-Aware Label Aggregation for LLM-as-a-Judge via Ising Models

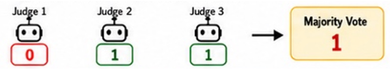
Krishnakumar Balasubramanian[1,2] • Aleksandr Podkopaev[1] • Shiva Kasiviswanathan[1]
[1]Amazon Web Services, [2]UC Davis

**LLM judges are correlated. Ignoring dependence can be overconfident and wrong.** Ising models explicitly capture judge dependence and achieve better aggregation.

## 1. WHY MAJORITY VOTE CAN FAIL?

### Majority voting is popular

Multiple LLM judges are commonly aggregated using majority voting because it is simple and scalable

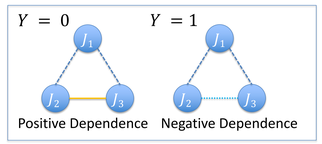



### Hidden Assumption: Independence

Majority vote implicitly treats judges as independent sources of evidence
However, LLM judges often share


### A Simple Counterexample

1. Assume three judges $(J_1, J_2, J_3 \in \{0,1\})$ are sampled from a class-dependent Ising model
   Dependence structure:
   When $Y = 0$, $J_2$ and $J_3$ tend to agree
   When $Y = 1$, $J_2$ and $J_3$ tend to disagree




2. Suppose we observe votes $\boldsymbol{J} = (J_1, J_2, J_3) = (0,1,1)$
3. Bayes-optimal prediction
   The observed vote pattern is substantially more likely under $Y = 0$ than under $Y = 1$,
   $P(Y = 0 \mid J_1 = 0, J_2 = 1, J_3 = 1) > P(Y = 1 \mid J_1 = 0, J_2 = 1, J_3 = 1)$
   **Bayes-optimal prediction $Y = 0$**
4. Majority vote prediction
   Two votes for class 1 → **Majority prediction $Y = 1$**

| Key Insight | Dependence is information | The agreement structure among judges can be more informative than the vote count itself | This intuition can be made precise |
|---|---|---|---|

### Average Ising Coupling Within Model Families

Same-provider models share more residual dependence than cross-provider pairs, reflecting shared training data, RLHF procedures, and architectural choices within each model family

| Dataset | Anthropic–Anthropic | Meta–Meta | DeepSeek–DeepSeek | Cross-family |
|---|---|---|---|---|
| Relevance | 0.400 | 0.397 | 0.475 | 0.375 |
| Toxicity | 0.441 | 0.488 | 0.572 | 0.296 |
| Summarization | 0.316 | 0.152 | 0.098 | 0.111 |

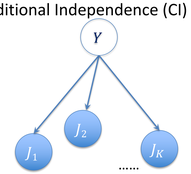
Within-family couplings consistently exceed cross-family couplings
→ LLM judges from the same provider are NOT independent

## 2. DEPENDENCE AWARE AGGREGATION

### Graphical Models

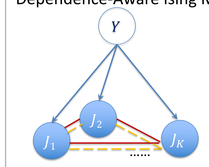
$\boldsymbol{J} = (J_1, \ldots, J_K)$ denote the judge labels
$P(\boldsymbol{J} \mid Y = y) \propto exp(\boldsymbol{J}^T h^{(y)} + \boldsymbol{J}^T W^{(y)} \boldsymbol{J})$

Conditional Independence (CI)




Judges are independent given $Y$
$P(\boldsymbol{J} \mid Y) = \prod_{i=1}^{m} P(J_i \mid Y)$

Dependence-Aware Ising Model

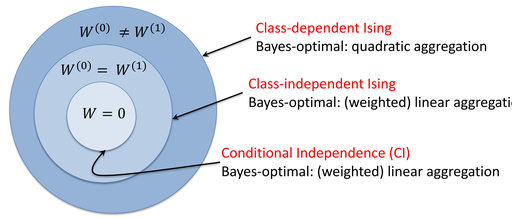



$W^{(1)}$ when $Y = 1$
$W^{(0)}$ when $Y = 0$

$h^{(y)} \in \mathcal{R}^K$: vector of class-$y$ local fields capturing per-judge bias/strength

$W^{(y)} \in \mathcal{R}^{K \times K}$: class-$y$ coupling matrix encoding pairwise dependencies

### Model Hierarchy (Set-Inclusion)




### Main Theoretical Result

Theorem: Let $\pi = P(Y = 1)$ denote the class prior and let $R(g) = P(g(\boldsymbol{J}) \neq Y)$ denote the classification risk of the binary aggregator $g$.

There exist class-conditional **Curie–Weiss Ising** models where every judge has identical one-dimensional marginals under both labels:

$$P(J_i = 1 \mid Y = 0) = P(J_i = 1 \mid Y = 1) = \frac{1}{2} \text{ for all } i \in [m]$$

where the conditional-independence aggregator cannot extract any information from the individual judge votes and attains risk

$$R\left(g_K^{\mathrm{CI}}\right) = \min\{\pi, 1 - \pi\}.$$

In contrast, the Bayes-optimal aggregator achieves

$$R(g_K^*) \to 0$$

as the number of judges $K$ grows. Consequently,

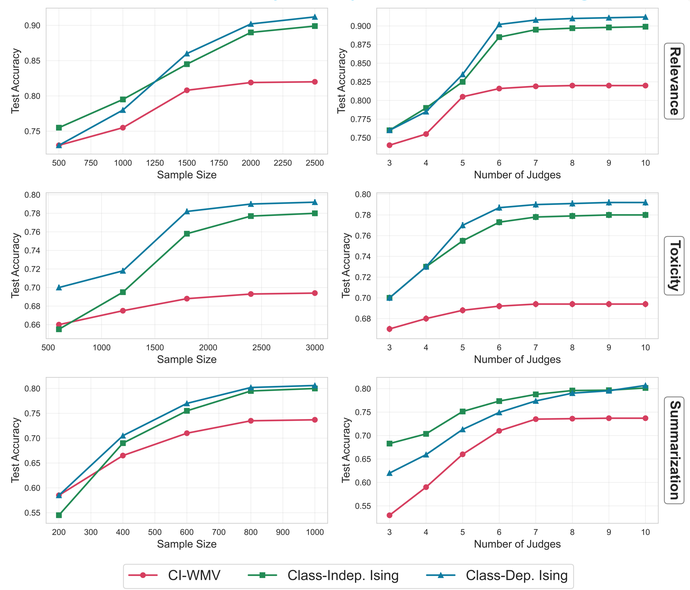
$$\lim_{K \to \infty} R(g_K^{\mathrm{CI}}) - R(g_K^*) \to \min\{\pi, 1 - \pi\} > 0.$$

Interpretation: Even with infinitely many judges, CI fails because the label is encoded only in the dependence structure, not in individual judge marginals. Therefore, a non-vanishing gap separates dependence-aware and dependence-ignorant aggregation.

Extension: The above separation persists even when each individual judge is informative (better than random).

## 3. EXPERIMENTAL RESULTS

### Test Accuracy vs. Sample Size and Number of Judges




Judge Models (K = 10): Anthropic (4), OpenAI (2), Meta (3), DeepSeek (2)
Tasks: Relevance (Web search result quality rating), Toxicity (Content safety classification), Summarization (Summary quality assessment)

**Learning:** Model parameters are estimated using a generalized EM algorithm that alternates between inferring latent labels (E-step) and updating class-conditional Ising parameters via weighted pseudo-likelihood (M-step).

### Average Test Accuracy (Higher is better)

| Dataset | CI-MVW | Class-Indep. Ising | Class-Dep. Ising |
|---|---|---|---|
| Relevance | 0.820 | 0.899 | 0.912 |
| Toxicity | 0.694 | 0.780 | 0.792 |
| Summarization | 0.737 | 0.806 | 0.801 |

### Key Experimental Findings

- ✅ Dependence-aware aggregation consistently outperforms CI baselines
- ✅ Gains increase with more judges and more training data
- ✅ Class-Dependent Ising achieves the strongest performance when sufficient data are available

### Practical Takeaway

**LLM judges are rarely independent.** Shared training data, architectures, prompts, and failure modes induce correlations that can bias simple aggregation rules. **Modeling these dependencies leads to more accurate label aggregation.**