Don't Always Pick the Highest-Performing Model: An Information Theoretic View of LLM Ensemble Selection

Yigit Turkmen [1] **Baturalp Buyukates** [2] **Melih Bastopcu** [1]

Abstract

Large language models (LLMs) are often ensembled together to improve overall reliability and robustness, but in practice models are strongly correlated. This raises a fundamental question: which models should be selected when forming an LLM ensemble? We formulate budgeted ensemble selection as maximizing the mutual information between the true label and predictions of the selected models. Furthermore, to explain why performance can saturate even with many models, we model the correlated errors of the models using Gaussian-copula and show an information-theoretic error floor for the performance of the ensemble. Motivated by these, we propose a simple greedy mutual-information selection algorithm that estimates the required information terms directly from data and iteratively builds an ensemble under a query budget. We test our approach in two question answering datasets and one binary sentiment classification dataset: MEDMCQA, MMLU, and IMDB movie reviews. Across all datasets, we observe that our method consistently outperforms strong baselines under the same query budget.

1. Introduction

Large language models (LLMs) are increasingly used as general purpose decision engines in various tasks (Minaee et al., 2025), yet their outputs can be unstable: a single model may fail on particular prompts due to reasoning gaps (Huang & Chang, 2023), ambiguity (Lu et al., 2022) or distribution shift (Yang et al., 2023). A standard approach in machine learning is to ensemble multiple predictors to improve robustness and accuracy of the overall system at the expense of extra latency and monitory costs (Breiman, 1996; Dietterich, 2000; Chen et al., 2025). For modern LLM systems, this cost is often restrictive because real deployments typically cannot query a full pool of models for every request. In this paper, we develop our framework for general multiple label (M-ary) classification under an arbitrary class prior, and we test it empirically in the binary case where statements are classified as true or false (positive or negative) using an ensemble of LLM responses, as shown in Figure 1. The binary instantiation is not a limitation of the theory but a deliberate choice for empirical clarity: it captures the canonical LLM-as-judge deployment shape i.e. verification, fact-checking, sentiment, safety, claim entailment, reward modeling. To this end, we study the budgeted ensemble selection problem: given a pool of K LLMs and a per-query budget of $k < K$ calls, we want to form a subset S of size k that minimizes the final decision error under an optimal (or near-optimal) aggregator. The key difficulty is that LLM errors are strongly *correlated*. Models from the same family tend to share training data, architectures, and failure modes, so adding another high-accuracy model may contribute little new information or may even degrade performance if it reinforces the same mistakes (Cemri et al., 2025; Kim et al., 2025), a phenomenon illustrated in Figure 2. This makes the common selecting "Top-k by accuracy" heuristic unreliable in practice.

To address this unreliability and the resulting subset-selection problem, we take an information-theoretic view. We treat each model output as a noisy observation of a binary ground-truth label Y and use mutual information to quantify how informative a selected subset is about Y. Our analysis goes from the ideal to the realistic: when model errors are effectively independent, we show that the best use of a budget of k queries is simply picking the k most accurate models, providing a clean rationale for the Top-k baseline. In the correlated regime, however, the value of adding a model depends not only on its accuracy but also on how much *new information* it contributes relative to the current subset; we formalize this trade-off via a decomposition that separates prediction redundancy from structured dependence in error

This work was supported by Tubitak 2232-B program (Project No:124C533).

[1]Department of Electrical and Electronics Engineering, Bilkent University, Ankara, Turkey [2]School of Computer Science, University of Birmingham, Birmingham, UK. Correspondence to: Yigit Turkmen <yigit.turkmen@ug.bilkent.edu.tr>, Melih Bastopcu <bastopcu@bilkent.edu.tr>, Baturalp Buyukates <b.buyukates@bham.ac.uk>.

Accepted to the ICML 2026 Workshop on Statistical Frameworks for Uncertainty in Agentic Systems, Seoul, South Korea, 2026.
Copyright 2026 by the author(s).

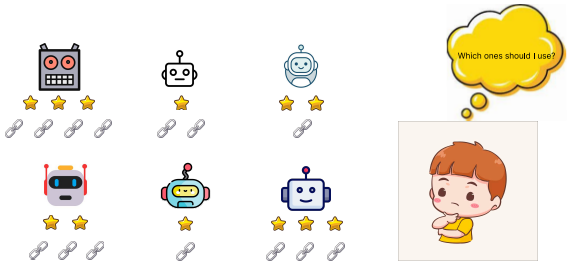


Figure 1. Candidate LLM pool for ensembling: stars indicate each model's accuracy, while chain links denote pairwise correlation (shared error patterns) with other models.

(a) Diverse Ensemble

Q: Axonal transport is: Is the answer 'Antegrade and retrograde'?
Ground Truth: `True`

Model	Acc.	Ans.
GPT-4o	80%	T ✗
Gemini-2.5-F.	73%	F ✗
Claude-3.5-H.	71%	T ✗
Llama-3.1-8B	66%	T ✗
Avg.	**72%**	**3/4** ✓

(b) Top-k Accuracy

Q: Axonal transport is: Is the answer 'Antegrade and retrograde'?
Ground Truth: `True`

Model	Acc.	Ans.
GPT-5-Chat	83%	F ✗
GPT-4.1	82%	F ✗
GPT-4o	80%	T ✗
GPT-4.1-mini	78%	F ✗
Avg.	**81%**	**1/4** ✗

Figure 2. Example from MEDMCQA. (a) A diverse ensemble with **lower average accuracy** (72%) answers correctly by combining models from different families. (b) Selecting the ensemble with strongest models with **higher average accuracy** (81%) fails because almost all GPT models share the same error pattern for this example. Please note that the original multiple-choice question is converted to binary format: for each candidate answer, we ask "Is the answer '[candidate]'?" with ground truth $Y = +1$ (correct) or $Y = -1$ (incorrect).

patterns, motivating a *greedy mutual-information selection* strategy. To model such dependencies and study scaling limits, we use a *Gaussian-copula* to represent the joint law of LLM responses (equivalently, correlated error events) and characterize an explicit *saturation floor* under a latent-factor interpretation: beyond a point, adding additional models cannot eliminate the uncertainty induced by shared difficulty. This difficulty-aware perspective helps explain why greedy mutual information selection yields strong gains in practice while still exhibiting diminishing returns in highly correlated ensembles.

Contributions: We make the following contributions:

- We represent LLM decisions using a Gaussian-copula, which allows us to capture both the individual accuracies and the correlations among models. Our numerical results show that the error behavior of real models can be effectively captured by the simulated Gaussian-copula model.
- We first consider independent LLM decisions. For this special case, we show in Theorem 3.1 that if LLMs were to make independent errors, the optimal model selection would be trivially selecting the most accurate models. The breakdown of Top-k selection in practice therefore arises solely from the underlying correlation structure.
- Since identifying the LLM subset that minimizes the probability of error is a combinatorial problem, we instead propose a greedy selection strategy that maximizes mutual information gain. In Theorem 3.3, we characterize the underlying principle of this selection rule and establish its connection to the Maximum Relevance–Minimum Redundancy (mRMR) principle from feature selection (Peng et al., 2005).
- We then characterize a non-vanishing error floor under correlated errors (Theorem 3.4), with closed forms in the balanced M-ary and imbalanced binary cases and an implicit characterization in the fully general setting. This result explains why ensembling an increasing number of correlated models may fail to further reduce the error probability.
- Finally, we evaluate our approach on two question-answering benchmarks and a binary sentiment classification task that are MEDMCQA, MMLU, and IMDB movie reviews – instantiated in the binary balanced setting – and demonstrate consistent improvements over strong baselines such as the "Top-k by accuracy" under identical query budgets.

Due to space constraints, a comprehensive review of related work is provided in Appendix A

2. System Model and Problem Formulation

In this work, we consider a multiple label (M-ary) classification problem carried out by a group of LLMs, where each incoming task involves selecting one label from $[M] \triangleq \{1, \ldots, M\}$. We model the correct label with a random variable $Y \in [M]$ drawn from a class prior $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_M)$ with $\pi_y > 0$ for all y and $\sum_{y=1}^{M} \pi_y = 1$, allowing arbitrary (possibly imbalanced) priors. In order to perform these tasks, we have access to an ensemble of K LLMs, indexed by the set $[K] \triangleq \{1, \ldots, K\}$. For a given input instance, model j produces an M-ary prediction $X_j \in [M]$.

We model each LLM as an independent M-ary symmetric channel (MSC) with error rate $\epsilon_j \in [0, (M-1)/M)$:

$$\mathbb{P}(X_j = y' \mid Y = y) = \begin{cases} 1 - \epsilon_j, & y' = y, \\ \epsilon_j/(M-1), & y' \neq y, \end{cases} \tag{1}$$

with X_j's conditionally independent across j given Y. The error indicator for LLM j is $E_j \triangleq \mathbb{1}(X_j \neq Y)$, with

marginal error probability $\mathbb{P}(E_j = 1) = \epsilon_j$ and accuracy $\alpha_j = 1 - \epsilon_j$. We denote the vector of responses for a subset $S \subseteq [K]$ as $\mathbf{X_S} \triangleq (X_j)_{j \in S}$.

Our experiments instantiate the binary balanced case ($M = 2$ with $\mathbb{P}(Y = +1) = \mathbb{P}(Y = -1) = 1/2$), where $Y = +1$ denotes a true statement and $Y = -1$ a false one. In this case (1) reduces to a binary symmetric channel (BSC) and the response is given by the sign-flip $X_j = Y \cdot (-1)^{E_j}$.

2.1. Modeling Latent Errors with Gaussian-Copula

In general, when responses are collected from multiple LLMs, their decisions may not be independent. Models trained on similar data or architectures tend to exhibit correlated errors. To capture such a complex dependency structure of LLM errors, in this work, we model the joint error distribution using a Gaussian-copula (Nelsen, 2007). With this goal, we define $\mathbf{Z} = (Z_1, \ldots, Z_K)$ to be a latent random vector following a multivariate normal distribution $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ where $\mathbf{\Sigma} \in \mathbb{R}^{K \times K}$ is a correlation matrix with unit diagonal entries and off-diagonal entries given by ρ_{ij} capturing the pairwise latent correlation between models i and j where $i \neq j$. In the case of equicorrelated ensembles where all the correlations of the models are equal to each other, i.e., $\rho_{ij} = \rho$ for all i and j with $i \neq j$, this dependence structure can be equivalently represented by a linear factor model:

$$Z_j = \sqrt{\rho}\, U + \sqrt{1 - \rho}\, \xi_j, \tag{2}$$

where $U \sim \mathcal{N}(0, 1)$ is a common latent factor shared by all models and $\xi_j \sim \mathcal{N}(0, 1)$ are independent individual noise terms. Then, the binary error indicator E_j can be generated by thresholding the latent variable Z_j as

$$E_j = \mathbb{1}(Z_j < \tau_j), \tag{3}$$

where the threshold $\tau_j = \Phi^{-1}(\epsilon_j)$ is determined by the marginal error probability $\epsilon_j = \mathbb{P}(E_j = 1)$ of model j, and $\Phi(\cdot)$ is the standard normal cumulative distribution function and its inverse is denoted by $\Phi^{-1}(\cdot)$. Note that the Gaussian-copula models the dependence structure of the binary error events $(E_1, \ldots, E_K)$, which is alphabet-agnostic: the same construction applies in both the binary and M-ary settings, with only the coupling between E_j and the observed output X_j depending on M.

By modeling the error via Gaussian-copula, we are able to: $i)$ choose the variables $\{\tau_j\}_{j=1}^{K}$that uniquely determine the individual error probability ϵ_j for each model, independent of the others, and $ii)$ model the dependence structure through the covariance matrix $\mathbf{\Sigma}$ that governs the higher-order error correlations. In other words, the coefficient ρ_{ij} captures the tendency of models to exhibit correlated errors on identical inputs. Using a Gaussian-copula, we can explicitly decouple the marginal statistics from the dependence structure.

2.2. Ensemble Aggregation and Selection

When the joint error distribution of the LLMs is known, the decision rule that minimizes the error probability under prior $\boldsymbol{\pi}$ is the Maximum A Posteriori (MAP) estimator (Hajek, 2020). Thus, for a given selected subset of models S and their observed outputs $\mathbf{X_S} = \mathbf{x_S}$, the MAP rule is given by

$$\widehat{Y}_{\mathrm{MAP}}(\mathbf{x_S}) = \arg\max_{y \in [M]} \pi_y \, \mathbb{P}(\mathbf{X_S} = \mathbf{x_S} \mid Y = y). \tag{4}$$

The performance of any subset S is measured by its error probability given by $P_e(S) \triangleq \mathbb{P}(\widehat{Y}_{\mathrm{MAP}}(\mathbf{X_S}) \neq Y)$. In the binary balanced instantiation used in our experiments ($M = 2$, $\pi_{+1} = \pi_{-1} = 1/2$), the prior weights drop out and $P_e(S)$ reduces to the standard probability-of-error metric. We implement Algorithm 1 for the MAP aggregation in this case.

2.3. Problem Formulation

Due to inference costs and latency constraints, collecting responses from all LLMs may not be feasible for each query. Thus, for a given query, our goal is to select the subset $S_k^\star$ of size k minimizing the probability of error:

$$S_k^\star \in \arg\min_{S \subseteq [K], |S| = k} P_e(S). \tag{5}$$

Finding $S_k^\star$ requires choosing k LLMs from the K available models, which is combinatorially hard and depends on the joint error distribution $\mathbb{P}(E_1, \ldots, E_K)$. We capture this distribution via the Gaussian-copula latent structure of Section 2.1: there exists $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \Sigma)$ and thresholds $\{\tau_j\}$ such that $E_j = \mathbb{1}(Z_j < \tau_j)$, with Σ a correlation matrix capturing cross-model error dependence. The independent MSC regime corresponds to $\Sigma = \mathbf{I}$ (i.e., $\rho = 0$) and is studied in Theorem 3.1; the equicorrelated regime $\rho_{ij} = \rho$ enables the closed-form saturation limits of Theorem 3.4. The binary symmetric ensemble used in our experiments is the further special case $M = 2$.

3. Information-Theoretic Analysis of Ensemble Selection

The central challenge in ensemble selection is that the error-optimal subset $S^\star$ is combinatorially hard to find. Practical systems typically rely on heuristics like selecting top-k accurate models. In this section, we analyze the optimality of this baseline and contrast it with greedy mutual information maximization, which serves as the natural information-theoretic alternative. We first prove that they are equivalent in the independent case, then derive the precise information-theoretic penalty introduced by correlations, and finally establish a fundamental limit on the achievable error. Moreover, we present Theorem F.1 in Appendix D which further explains why greedy approaches yet can saturate under latent difficulty.

3.1. Optimal Subset Selection Under Independence

We begin by analyzing the baseline case where model errors are statistically independent. This serves as a control setting to demonstrate that accuracy-based and information-theoretic objectives coincide in the absence of response correlations. Before stating the following theorem, we note that our system resembles a well-known communication channel consisting of K independent M-ary symmetric channels (MSCs), each of which corresponds to an LLM in our system, while the M-ary information source corresponds to the underlying classification task. Thus, our goal can be translated into selecting the best subset of k MSCs so as to minimize the probability of error in this communication system.

Theorem 3.1 (Alignment of Error and Mutual Information)**.** *Consider an ensemble of independent MSCs where channel j has error rate $\epsilon_j < (M-1)/M$, under an arbitrary class prior $\boldsymbol{\pi}$ with full support on $[M]$. Let the channels be indexed by accuracy such that $\epsilon_1 \leq \epsilon_2 \leq \ldots \leq \epsilon_K$. Let $H_k = \{1, \ldots, k\}$ be the set of top-k most accurate channels. Let $S_k^{MI} \in \operatorname{argmax}_{|S|=k} I(Y; \mathbf{X_S})$ denote the subset of size k that maximizes the mutual information between Y and $\mathbf{X_S}$. For any subset $S \subseteq [K]$ with $|S| = k$, the top-k set H_k dominates S both in terms of maximizing the mutual information and minimizing the Bayes risk. Thus, we have*

$$I(Y; \mathbf{X_{H_k}}) \geq I(Y; \mathbf{X_S}), \tag{6}$$

$$\mathbb{P}_e(H_k) \leq \mathbb{P}_e(S). \tag{7}$$

Consequently, when the channels are independent and symmetric, the error-optimal subset $S^\star$ and the mutual-information optimal subset S_k^{MI} coincide with H_k, regardless of the label cardinality M and the class prior $\boldsymbol{\pi}$.

The proof of Theorem 3.1 is provided in Appendix B and proceeds via an M-ary stochastic-degradation argument that recovers the binary case ($M=2$) and the balanced prior as special instances. In the following remark, we state how the result of Theorem 3.1 applies to our problem.

Remark 3.2. Theorem 3.1 confirms that if LLMs make independent errors, the optimum selection of LLMs would be trivial: simply picking the most accurate k models is optimal, regardless of label cardinality or class imbalance. The failure of Top-k selection in practice is therefore purely a consequence of the correlation structure derived in the ensuing analysis not a consequence of the binary or balanced framing of our experiments.

3.2. Decomposition of Correlated Ensembles

When models are correlated, the independence assumption in Theorem 3.1 breaks. To quantify exactly how correlation alters the selection objective, we derive a decomposition of the marginal information gain.

Theorem 3.3 (Accuracy-Redundancy-Error Decomposition)**.** *For LLM outputs $X_j \in [M]$ and true label $Y \in [M]$ drawn from an arbitrary prior $\boldsymbol{\pi}$, let $E_j = \mathbb{1}(X_j \neq Y)$ be the error indicator for model j, and we define the marginal information gain of adding model j to subset S as*

$$\Delta(j \mid S) \triangleq I(Y; X_j \mid \mathbf{X_S}). \tag{8}$$

Then, $\Delta(j \mid S)$ admits the following decomposition

$$\Delta(j \mid S) = I(Y; X_j) - I(X_j; \mathbf{X_S}) + I(E_j; E_S) + \Lambda_j(S), \tag{9}$$

where $E_S = \{E_j \mid j \in S\}$ and we can define the last term in (9) as the label-dependence correction term

$$\Lambda_j(S) \triangleq I(E_j; Y \mid E_S) - I(E_j; Y). \tag{10}$$

Under the label-invariant error assumption, i.e., $(E_1, \ldots, E_m) \perp Y$, we have $\Lambda_j(S) = 0$ for all j and S, the decomposition for $\Delta(j \mid S)$ reduces to

$$\Delta(j \mid S) = I(Y; X_j) - I(X_j; \mathbf{X_S}) + I(E_j; E_S). \tag{11}$$

The proof of Theorem 3.3 is provided in Appendix C. Under the label-invariant error assumption, this decomposition highlights two opposing forces: $i)$ Prediction Redundancy $I(X_j; \mathbf{X_S})$: Appears in (11) with a negative sign. This term penalizes models that mimic the ensemble's output. $ii)$ Error Correlation $I(E_j; E_S)$: Appears in (11) with a positive sign. Counter-intuitively, correlated errors can be beneficial when structured because knowing the error pattern of S helps to decode Y.

Standard diversity heuristics often mirror the mRMR principle from feature selection (Peng et al., 2005), favoring high relevance while discouraging redundancy among predictors. Theorem 3.3 shows that under the label-invariant error assumption $I(E; Y) = 0$, and the general M-ary error representation $E_j = \mathbb{1}(X_j \neq Y)$, the marginal information gain admits an additional term, $I(E_j; E_S)$, alongside the usual relevance and redundancy terms. Consequently, mRMR-style criteria based solely on $(I(Y; X_j),\ I(X_j; \mathbf{X_S}))$ are not guaranteed to align with mutual information-optimal subset selection in correlated ensembles, since they do not account for structured dependence among model *errors*.

Thus, in Theorem 3.3, we propose an information-theoretic criterion for deciding whether to include a model in a subset when models are correlated. As seen in (11), this criterion resembles the well-known mRMR-style feature selection problem (Peng et al., 2005), augmented with an additional term, $I(E_j; E_S)$, which explicitly accounts for error correlations among models.

3.3. Asymptotic Saturation Limit

In Theorem 3.1, we show that under independent error assumptions, subset selection reduces to selecting the most

accurate models. The decomposition in Theorem 3.3 then explains how this picture changes under dependence, where correlation introduces additional structure beyond simple redundancy. We now address the central question left open by these results: *Can ensemble error vanish as the ensemble size grows when model errors are correlated?* Under the Gaussian-copula model with uniform pairwise correlation, we show that the answer to this is generally no: as K grows, the probability of error converges to a non-zero floor determined by the correlation structure and the label cardinality, yielding a fundamental saturation limit.

Theorem 3.4 (MAP Information Saturation)**.** *Consider an infinite ensemble generated by the Gaussian-copula model with uniform pairwise correlation $\rho > 0$ and the marginal accuracy $\alpha \triangleq 1-\epsilon > 1/M$, under a balanced M-ary prior $\boldsymbol{\pi} = (1/M, \ldots, 1/M)$. Let $\widehat{Y}_{\mathrm{MAP}}$ be the optimal MAP estimator given the full ensemble outputs $\mathbf{X}_{[\mathbf{K}]}$. The Bayes risk is lower-bounded by:*

$$\lim_{K\to\infty} \mathbb{P}(\widehat{Y}_{\mathrm{MAP}} \neq Y) = \Phi\left(\frac{\Phi^{-1}(1-\alpha) - \sqrt{1-\rho}\,\Phi^{-1}((M-1)/M)}{\sqrt{\rho}}\right) > 0. \tag{12}$$

The proof of Theorem 3.4 is provided in Appendix D. Theorem 3.4 establishes that the error floor observed in correlated ensembles is not negligible. It arises not from the sub-optimality of the aggregation rule (since the MAP is optimal), but from the information-theoretic capacity limit imposed by the latent common noise U. This implies that scaling the ensemble size K cannot overcome structural correlations; performance improvements must instead come from reducing the effective ρ via subset selection. The binary balanced case used in our experiments is recovered by setting $M = 2$, in which $\Phi^{-1}(1/2) = 0$ and the bound reduces to the closed form $\Phi(\Phi^{-1}(1-\alpha)/\sqrt{\rho})$.

We next show that the saturation floor is robust to the class prior: imbalanced priors do not change the asymptotic Bayes risk, only the rate at which it is approached.

Corollary 3.5 (Prior-Invariance of the Saturation Floor)**.** *Under the same Gaussian-copula model as in Theorem 3.4, but with an arbitrary class prior $\boldsymbol{\pi}$ on $[M]$ satisfying $\pi_y > 0$ for all y, the asymptotic Bayes risk satisfies*

$$\lim_{K\to\infty} \mathbb{P}(\widehat{Y}_{\mathrm{MAP}} \neq Y) = \Phi\left(\frac{\Phi^{-1}(1-\alpha) - \sqrt{1-\rho}\,\Phi^{-1}((M-1)/M)}{\sqrt{\rho}}\right), \tag{13}$$

which is identical to the balanced-prior expression in Theorem 3.4.

The proof of Corollary 3.5 is provided in Appendix E. Intuitively, the MAP rule under prior $\boldsymbol{\pi}$ adds a finite log-prior offset $\log \pi_y$ to the per-hypothesis log-likelihood, while the log-likelihood signal from K conditionally independent ensemble outputs grows linearly in K. As $K \to \infty$, the finite offset is dominated by the growing likelihood signal, and the asymptotic error event coincides with the balanced case. We note that imbalanced priors can affect the *rate* of convergence to the floor – more skewed priors require more samples to overcome the prior bias – but not the floor itself.

Together, Theorem 3.4 and Corollary 3.5 show that the saturation phenomenon depends only on the latent error structure (ρ), the marginal accuracy (α), and the label cardinality (M), and is robust to both class imbalance and the binary framing of our experiments.

4. Proposed Selection Algorithm: Greedy Mutual Information (Greedy MI)

Motivated by the decomposition in Theorem 3.3, we propose an efficient greedy strategy to approximate the optimal subset $S^\star$. Since the optimization problem in (5) is NP-hard, we adopt a stepwise maximization approach. This greedy strategy is computationally tractable and is roughly related to relevance–redundancy feature selection. Crucially, in correlated ensembles the marginal gain contains additional dependence terms, shown in (11), so the classical mRMR intuition does not directly apply to our problem. The selection criterion is alphabet- and prior-agnostic, applying directly in both the binary balanced setting used in our experiments and the general M-ary imbalanced framework introduced in Section 2.

4.1. Selection Criterion

We construct the subset S iteratively. Let S_{k-1} be the set of $k-1$ models selected so far. At step k, we select the model $j^\star$ that maximizes the marginal information gain:

$$j^\star = \operatorname*{argmax}_{j \in [K] \setminus S_{k-1}} \Delta(j \mid S_{k-1}). \tag{14}$$

By using Theorem 3.3, we explicitly calculate this score by estimating the three components using empirical mutual information obtained from the training set (introduced in Section 5); this split is used only to estimate statistics and to fit the empirical MAP estimator, and does not involve training any LLM parameters. We treat the model outputs as discrete random variables and estimate their joint distributions via frequency counts. Then, we calculate the estimated marginal information gain of model j by

$$\hat{\Delta}(j \mid S_{k-1}) = \hat{I}(Y; X_j) - \hat{I}(X_j; X_{S_{k-1}}) + \hat{I}(E_j; E_{S_{k-1}}). \tag{15}$$

This update rule balances the individual accuracy against prediction redundancy while explicitly rewarding models that introduce structured, predictable error patterns. Finally,

Algorithm 1 Greedy MI Selection

Input: Estimate the dist. of set outputs $X_{[K]}$, labels Y, target size k
Initialize $S \leftarrow \emptyset$
for $t = 1$ **to** k **do**
  Estimate each term in Equation (11) using Algorithm 2
  Compute score $\hat{\Delta}(j|S)$ for all $j \notin S$ using (11)
  $j^\star \leftarrow \text{argmax}_j \hat{\Delta}(j|S)$
  $S \leftarrow S \cup \{j^\star\}$
end for
Output: Selected subset $S_k^* = S$

we provide Algorithm 1 to summarize the steps for the Greedy MI selection proposed in this work. Note that, in practice, we compute the full conditional mutual information $\Delta(j \mid S) = I(Y; X_j \mid X_S)$ directly to account for potential label-dependent error structure in the data.

5. Experimental Results

We instantiate the framework in the binary balanced setting and evaluate Greedy MI against six baselines on three datasets. Binary classification covers the canonical LLM-as-judge deployment shape (verification, fact-checking, sentiment) as mentioned in the introduction and it admits clean copula validation via tetrachoric correlation;therefore, the binary instantiation is a deliberate part of the general M-ary imbalanced framework developed in Section 2, not a limitation.

Baselines. We compare Greedy MI to six selection strategies: *Top-k* (selects the most accurate models); *Term 1* (ranks by $\hat{I}(Y; X_j)$, the first term of (15)); *mRMR* (uses Terms 1+2, excluding the error-correlation term); *TMC-Shapley* (Ghorbani & Zou, 2019) with 5000 Monte-Carlo permutations and MAP utility; *LogReg Forward Selection* with 5-fold inner CV; and *LLM-TOPLA* (Tekin et al., 2024), which combines focal diversity with ensemble accuracy. All methods use the MAP aggregator at evaluation time. Implementation details are in Appendix G.

5.1. Experiments on MEDMCQA

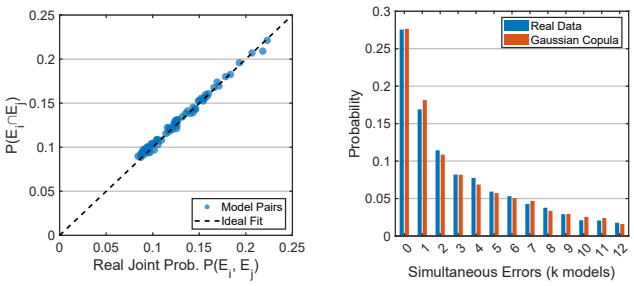
We first evaluate on MEDMCQA (Pal et al., 2022), a multiple-choice medical QA benchmark whose 4183-sample validation set is well suited to our setting because medical queries often induce shared failure modes across LLMs.

Binary conversion. Each multiple-choice sample is converted into one positive query ("Question: [Q]. Is the answer '[Correct]'?", $Y = +1$) and one negative query with a randomly chosen incorrect option ($Y = -1$), yielding 8366 balanced binary samples.

Table 1. Selected models on MEDMCQA at representative ensemble sizes (temp = 0.7, run 1).

k	**Greedy MI (acc.)**	**mRMR (acc.)**
1	GPT-5-chat (83.1%)	GPT-5-chat (83.1%)
2	+ GPT-4.1 (82.2%)	+ LLaMA-3.1-8b (66.0%)
3	+ Qwen3-235b (77.4%)	+ GPT-4.1-nano (66.8%)
4	+ kimi-k2 (76.3%)	+ Claude-3.5-haiku (70.7%)
5	+ Gemini-2.5-flash (72.9%)	+ Gemini-2.5-flash (72.9%)

Models and Setup. We use 12 LLMs spanning OpenAI (GPT-4.1, GPT-4.1-mini, GPT-4.1-nano, GPT-4o, GPT-5-chat) (OpenAI, 2026), LLaMA-3.1 (Grattafiori et al., 2024), Qwen3-235b (Yang et al., 2025a), kimi-k2 (Team et al., 2025), Claude-3.5-haiku (Anthropic, 2024), Mistral-small-3.1/3.2 (Mistral AI, 2026), and Gemini-2.5-flash (Comanici et al., 2025), run at three temperatures (0.01, 0.3, 0.7) with two independent runs each. For each setting, we use an 80/20 split, with 80% used to estimate the information-theoretic quantities (no LLM training is involved) and 20% for evaluation, repeated over 5 random splits. Per-model accuracies, the 12×12 correlation matrix (average $\bar{\rho}=0.55$), and the copula validation across all temperature/run pairs are reported in Appendix K and Appendix H.3.



(a) *(b)*

Figure 3. Gaussian-copula validation on MEDMCQA (temp = 0.7, run 1). (a) Empirical vs. copula-modeled pairwise joint error probabilities $P(E_i \cap E_j)$ across all $\binom{12}{2} = 66$ pairs; tight clustering around the diagonal indicates strong second-order fit. (b) Distribution of simultaneous errors across models; copula (orange) closely matches the observed tail (blue), confirming higher-order fit beyond the fitted pairwise statistics.

Copula validation. Figure 3 shows two diagnostics for the temp = 0.7 run 1 split. The pairwise scatter (left) confirms second-order fit by construction; the simultaneous-error histogram (right) confirms higher-order fit on quantities not used during fitting. The same pattern holds across all temperature/run pairs (Appendix H.3), supporting the Gaussian-copula as an appropriate model for LLM error dependence.

Results. Figure 4 and Table 4 (Appendix K) show the mean test error across temperatures, runs, and splits for all seven methods. Greedy MI achieves its best performance of 16.3% at $k = 5$ versus 17.0% for Top-k, with consistent gains in

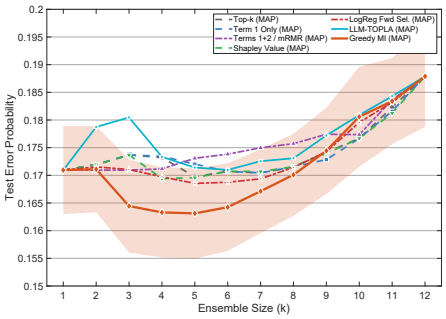


Figure 4. MEDMCQA test error vs. ensemble size, mean over temperatures, runs, and random splits. Shaded region is standard deviation around Greedy MI.

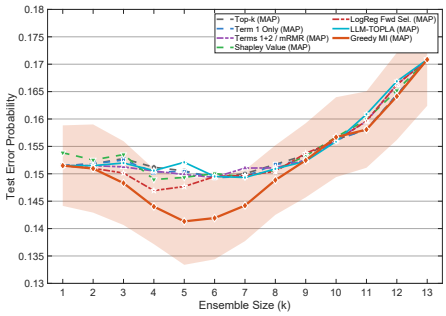


Figure 5. MMLU test error vs. ensemble size, mean over temperatures, runs, and random splits. Shaded region is standard deviation around Greedy MI.

the practically relevant mid-budget range $k \in [3, 7]$. One of the three baselines (TMC-Shapley, LogReg, LLM-TOPLA) sit between Top-k and Greedy MI but do not match Greedy MI in the mid-range; LogReg is the strongest of the three, closing some of the gap at $k \geq 5$ but never below Greedy MI.

Selection behavior. Table 1 contrasts Greedy MI and mRMR, the two methods whose qualitative behavior differs most. Both pick GPT-5-chat at $k=1$, but mRMR's aggressive redundancy penalty causes it to pick LLaMA-3.1-8b (66.0%) at $k=2$ and GPT-4.1-nano (66.8%, ranked 11th) at $k=3$, sacrificing accuracy for output diversity. Greedy MI instead picks GPT-4.1 then Qwen3-235b–staying within the top-5 by accuracy while diversifying across families. By $k=5$, Greedy MI's ensemble spans four families with all members above 72% accuracy; mRMR has already included two models below 71%. The correlation matrix in Figure 21 explains the dynamic: within-family correlations are high ($\rho \approx 0.7$–0.8, e.g. among OpenAI models) while cross-family correlations are moderate ($\rho \approx 0.4$–0.5), so cross-family diversity already reduces redundancy without forcing weak models into the ensemble. Top-k ignores correlation entirely and stacks four OpenAI models in succession, missing this structural opportunity.

Saturation For larger k, gains diminish across all methods, consistent with the saturation phenomenon characterized by Theorem 3.4. The MAP aggregator additionally requires estimating $\mathbb{P}(Y \mid \mathbf{X_S})$ over 2^k output patterns; with a fixed 80% training split, many patterns become rare or unseen at large k, inflating MAP variance and contributing to performance degradation by $k=12$. Full temperature/run grids and MV/W-MV ablations are in Appendix H.

5.2. Experiments on MMLU

We apply the same binary conversion, evaluation protocol, and 80/20 split to MMLU (Hendrycks et al., 2021), using 5000 randomly sampled questions (10000 binary queries) from the test split. The model pool is 13 LLMs: the MEDMCQA pool with GPT-5-chat replaced by GPT-4o-mini, LLaMA-3.3-70b added, and phi-4 (Abdin et al., 2024) added; details and the 13×13 correlation matrix are in Appendix K.

Table 2. Selected models on MMLU at representative ensemble sizes (temp $= 0.7$, run 2).

k	**Greedy MI (acc.)**	**mRMR (acc.)**
1	GPT-4.1 (85.0%)	GPT-4.1 (85.0%)
2	+ Qwen3-235b (82.2%)	+ LLaMA-3.3-70b (77.2%)
3	+ kimi-k2 (82.1%)	+ GPT-4.1-nano (69.2%)
4	+ Gemini-2.5-flash (79.8%)	+ Gemini-2.5-flash (79.8%)
5	+ GPT-4.1-mini (81.6%)	+ phi-4 (74.1%)
6	+ phi-4 (74.1%)	+ kimi-k2 (82.1%)

Figure 5 and Table 8 show that in the mid-budget regime where our gains are largest, Greedy MI remains consistently lower (e.g., $k = 4$: 0.144 vs. 0.151 for Top-k and LogReg, and 0.168 for LLM-TOPLA; $k = 6$: 0.142 vs. 0.145 for Top-k and 0.146 for TMC-Shapley). Top-k and Term 1 overlap nearly everywhere, confirming that accuracy and relevance rankings coincide when correlations are ignored. By $k=13$, all methods converge to ≈ 0.171, again driven by MAP estimator variance over the 2^k pattern space. The selection-behavior contrast in Table 2 mirrors MEDMCQA: mRMR diverges from Greedy MI at $k=2$ by picking LLaMA-3.3-70b (77.2%) over Qwen3-235b (82.2%), and selects GPT-4.1-nano (ranked 12/13) at $k=3$, while Greedy MI builds an ensemble entirely above 79% accuracy across four families by $k=5$. Full results, additional copula validation, and aggregation ablations are in Appendix I.

5.3. Experiments on IMDB Movie Reviews

We further provide experiments on the IMDB sentiment classification dataset (Maas et al., 2011), where models achieve uniformly high accuracies (91–96%) and exhibit very strong pairwise correlations ($\bar{\rho} = 0.90$). Due to this high correla-

tion, improvements from ensemble selection are modest, though Greedy MI still consistently outperforms baselines in the mid-budget range; full results are in Appendix J.

6. Conclusion

In this work, we studied the problem of ensembling responses from multiple large language models (LLMs) to improve accuracy under a fixed query budget. Working in the general M-ary classification setting under arbitrary class priors, we modeled LLM responses through an M-ary symmetric channel coupled to a Gaussian-copula on the latent error events, which effectively captures both individual model accuracies and inter-model correlations. We showed that when LLM decisions are independent, the optimal ensemble is obtained by selecting the most accurate models, corresponding to the classical "Top-k by accuracy" rule, regardless of label cardinality or class imbalance. However, in the presence of correlated models, accuracy alone is insufficient: even with an optimal estimator, correlations induce a non-vanishing ensemble error floor whose closed form depends only on the latent error structure, the marginal accuracy, and the label cardinality, and is robust to the class prior. Motivated by these results, we developed an information-theoretic framework, termed *Greedy MI*, that greedily constructs a near-optimal LLM ensemble under budget constraints. We tested our framework in the binary balanced setting and demonstrated its effectiveness by consistently outperforming six strong baselines across MEDMCQA, MMLU, and IMDB, with the largest gains in the practically relevant mid-budget regime.

7. Limitations and Discussion

Our theoretical framework covers M-ary classification under arbitrary class priors, but several limitations remain. First, the M-ary symmetric channel assumption distributes wrong-class probability uniformly over the remaining $M-1$ classes; real LLM confusion is often structured (e.g., consistent confusion of one specific class with another). Second, while our theory is fully general, our empirical evaluation uses only the binary balanced case; validating the framework on multi-class benchmarks and imbalanced-label scenarios is a natural next step. Third, the MAP aggregator requires estimating $\mathbb{P}(Y \mid \mathbf{X_S})$ over M^k output patterns, which introduces variance at large k when the training split is fixed; factorized aggregators (e.g., weighted log-odds, multinomial logit) may be needed for higher-cardinality settings, and separating selection-quality effects from aggregator-estimation effects in that regime is an open question. Finally, our saturation result (Theorem 3.4) is derived under uniform pairwise correlation; the non-equicorrelated case retains the saturation phenomenon but lacks a closed-form floor expression. Beyond these, the saturation effects revealed by our analysis indicate that further improvements in ensemble performance may come from complementing selection strategies with advances in model diversity and from larger training sets for estimating the information-theoretic quantities.

Impact Statement

This work advances the theoretical understanding of ensemble methods for large language models (LLMs) under practical budget constraints by explicitly accounting for correlations among model outputs. Although prior work has studied ensemble selection, existing solutions may not generalize well to all settings without a principled understanding of the problem's fundamental structure. By characterizing fundamental limits of ensemble performance and proposing principled selection strategies, our results can help practitioners design more reliable and cost-effective LLM systems, potentially reducing unnecessary computational and environmental costs associated with indiscriminate model querying. From an ethical perspective, the proposed framework does not introduce new data collection mechanisms or decision policies that directly affect individuals but rather provides analytical tools for improving existing aggregation methods. Overall, we expect the broader societal impact of this work to align with established implications of improving reliability and efficiency in machine learning systems.

References

Abdin, M., Aneja, J., Behl, H., Bubeck, S., Eldan, R., et al. Phi-4 technical report, 2024. URL https://arxiv.org/abs/2412.08905.

Ai, R., Pan, Y., Simchi-Levi, D., Tambe, M., and Xu, H. Beyond majority voting: LLM aggregation by leveraging higher-order information, 2025. URL https://arxiv.org/abs/2510.01499.

Anthropic. Model card addendum: Claude 3.5 Haiku and upgraded Claude 3.5 Sonnet. Technical report, Anthropic, 2024. URL https://assets.anthropic.com/m/1cd9d098ac3e6467/original/Claude-3-Model-Card-October-Addendum.pdf.

Anthropic. Anthropic models. https://platform.claude.com/docs/en/about-claude/models/overview, 2026. Accessed: 2026-01-01.

Bonett, D. G. and Price, R. M. Inferential methods for the tetrachoric correlation coefficient. *Journal of Educational and Behavioral Statistics*, 30(2):213–225, 2005. ISSN 10769986, 19351054. URL http://www.jstor.org/stable/3701350.

Breiman, L. Bagging predictors. *Mach. Learn.*, 24(2): 123–140, August 1996. ISSN 0885-6125. doi: 10.1023/A:1018054314350. URL https://doi.org/10.1023/A:1018054314350.

Brown, G., Pocock, A., Zhao, M.-J., and Luján, M. Conditional likelihood maximisation: A unifying framework for information theoretic feature selection. *Journal of Machine Learning Research*, 13(2):27–66, 2012. URL http://jmlr.org/papers/v13/brown12a.html.

Cemri, M., Pan, M. Z., Yang, S., Agrawal, L. A., Chopra, B., Tiwari, R., Keutzer, K., Parameswaran, A., Klein, D., Ramchandran, K., Zaharia, M., Gonzalez, J. E., and Stoica, I. Why do multi-agent LLM systems fail?, 2025. URL https://arxiv.org/abs/2503.13657.

Chen, L., Zaharia, M., and Zou, J. FrugalGPT: How to use large language models while reducing cost and improving performance. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=cSimKw5p6R. Featured Certification.

Chen, Z., Li, J., Chen, P., Li, Z., Sun, K., Luo, Y., Mao, Q., Li, M., Xiao, L., Yang, D., Ban, Y., Sun, H., and Yu, P. S. Harnessing multiple large language models: A survey on llm ensemble, 2025. URL https://arxiv.org/abs/2502.18036.

Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025. URL https://arxiv.org/abs/2507.06261.

Cover, T. M. and Thomas, J. A. *Elements of Information Theory 2nd Edition (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, July 2006. ISBN 0471241954.

Covert, I., Qiu, W., Lu, M., Kim, N., White, N., and Lee, S.-I. Learning to maximize mutual information for dynamic feature selection. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

DeepSeek-AI, Liu, A., Feng, B., Xue, B., Wang, B., et al. Deepseek-v3 technical report, 2025. URL https://arxiv.org/abs/2412.19437.

Dietterich, T. G. Ensemble methods in machine learning. In *Proceedings of the First International Workshop on Multiple Classifier Systems*, MCS '00, pp. 1–15, Berlin, Heidelberg, 2000. Springer-Verlag. ISBN 3540677046.

Elumar, E. C., Tekin, C., and Yagan, O. Cost-aware LLM-based online dataset annotation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=3AdTRYA2uJ.

Gadgil, S., Covert, I. C., and Lee, S.-I. Estimating conditional mutual information for dynamic feature selection. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Oju2Qu9jvn.

Ghorbani, A. and Zou, J. Data shapley: Equitable valuation of data for machine learning. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2242–2251. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/ghorbani19c.html.

Grattafiori, A., Dubey, A., Jauhri, A., et al. The Llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Hajek, B. Probability with engineering applications. https://courses.grainger.illinois.edu/ece313/fa2020/probabilityJan25.pdf, Jan. 2020. Lecture Notes, ECE 313, University of Illinois at Urbana–Champaign.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. Measuring massive multitask language understanding, 2021. URL https://arxiv.org/abs/2009.03300.

Huang, J. and Chang, K. C.-C. Towards reasoning in large language models: A survey. In Rogers, A., Boyd-Graber, J., and Okazaki, N. (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 1049–1065, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.67. URL https://aclanthology.org/2023.findings-acl.67/.

Jiang, D., Ren, X., and Lin, B. Y. LLM-Blender: ensembling large language models with pairwise comparison and generative fusion. In *Proceedings of the 61th Annual Meeting of the Association for Computational Linguistics (ACL 2023)*, 2023.

Kim, Y., Gu, K., Park, C., Park, C., Schmidgall, S., Heydari, A. A., Yan, Y., Zhang, Z., Zhuang, Y., Malhotra, M., Liang, P. P., Park, H. W., Yang, Y., Xu, X., Du, Y., Patel, S., Althoff, T., McDuff, D., and Liu, X. Towards a science of scaling agent systems, 2025. URL https://arxiv.org/abs/2512.08296.

Krause, A. and Golovin, D. Submodular function maximization. *Tractability*, 3(71-104):3, 2014.

Kruse, M., Afshar, M., Khatwani, S., Mayampurath, A., Chen, G., and Gao, Y. Simple yet effective: An information-theoretic approach to multi-LLM uncertainty quantification. In Christodoulopoulos, C., Chakraborty, T., Rose, C., and Peng, V. (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 30493–30504, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1551. URL https://aclanthology.org/2025.emnlp-main.1551/.

Li, D. X. On default correlation: A copula function approach. *The Journal of Fixed Income*, 9(4):43–54, 2000. doi: 10.3905/jfi.2000.319253.

Lu, Y., Bartolo, M., Moore, A., Riedel, S., and Stenetorp, P. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In Muresan, S., Nakov, P., and Villavicencio, A. (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8086–8098, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.556. URL https://aclanthology.org/2022.acl-long.556/.

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., and Potts, C. Learning word vectors for sentiment analysis. In Lin, D., Matsumoto, Y., and Mihalcea, R. (eds.), *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL https://aclanthology.org/P11-1015/.

Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., and Gao, J. Large language models: A survey, 2025. URL https://arxiv.org/abs/2402.06196.

Mistral AI. Mistral models. https://docs.mistral.ai/getting-started/models, 2026. Accessed: 2026-01-1.

Nelsen, R. *An Introduction to Copulas*. Springer Series in Statistics. Springer New York, 2007. ISBN 9780387286785. URL https://books.google.com.tr/books?id=yexFAAAAQBAJ.

OpenAI. Openai models. https://platform.openai.com/docs/models, 2026. Accessed: 2026-01-01.

Pal, A., Umapathi, L. K., and Sankarasubbu, M. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In Flores, G., Chen, G. H., Pollard, T., Ho, J. C., and Naumann, T. (eds.), *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pp. 248–260. PMLR, 07–08 Apr 2022. URL https://proceedings.mlr.press/v174/pal22a.html.

Pan, W., Xu, J., Chen, Q., Dong, J., Qin, L., Li, X., Yu, H., and Jia, X. Can LLMs refuse questions they do not know? measuring knowledge-aware refusal in factual tasks, 2025. URL https://arxiv.org/abs/2510.01782.

Peng, H., Long, F., and Ding, C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1226–1238, 2005. doi: 10.1109/TPAMI.2005.159.

Team, K., Bai, Y., Bao, Y., Chen, G., Chen, J., et al. Kimi K2: open agentic intelligence, 2025. URL https://arxiv.org/abs/2507.20534.

Tekin, S. F., Ilhan, F., Huang, T., Hu, S., and Liu, L. LLM-TOPLA: Efficient LLM ensemble by maximising diversity. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (eds.), *Findings of the Association for Computational Linguistics: EMNLP*

2024, pp. 11951–11966, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.698. URL https://aclanthology.org/2024.findings-emnlp.698/.

Turkmen, Y., Buyukates, B., and Bastopcu, M. Balancing information accuracy and response timeliness in networked LLMs, 2025. URL https://arxiv.org/abs/2508.02209.

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., and Zhou, D. Self-consistency improves chain of thought reasoning in language models, 2023. URL https://arxiv.org/abs/2203.11171.

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., et al. Qwen3 technical report, 2025a. URL https://arxiv.org/abs/2505.09388.

Yang, H., Li, M., Zhou, H., Xiao, Y., Fang, Q., Zhou, S., and Zhang, R. Large language model synergy for ensemble learning in medical question answering: Design and evaluation study. *Journal of Medical Internet Research*, 27:e70080, 2025b. doi: 10.2196/70080.

Yang, S., Chiang, W.-L., Zheng, L., Gonzalez, J. E., and Stoica, I. Rethinking benchmark and contamination for language models with rephrased samples, 2023. URL https://arxiv.org/abs/2311.04850.

Appendix

Table 3. Organization of the Appendix.

Section	Contents
A	Related Work
B	Proof of Theorem 3.1 (independence optimality)
C	Proof of Theorem 3.3 (MI decomposition)
D	Proof of Theorem 3.4 (saturation limit)
E	Proof of Corollary C.5 (Prior-Invariance of the Saturation Floor)
F	Difficulty-Aware Decomposition and the Price of Difficulty
G	Implementation Details and Algorithms
H	Additional Experiments on MEDMCQA
I	Additional Experiments on MMLU
J	Additional Experiments on IMDB
K	Supplementary Tables

A. Related Work

In this section, we provide related works on LLM ensembles, information theoretic feature selection problem, and Gaussian-copula models.

LLM Ensembles: The simplest approach to combining LLM outputs is majority voting, popularized by self-consistency (Wang et al., 2023) for single-model sampling and then extended to multi-model settings through weighted voting (Yang et al., 2025b; Jiang et al., 2023) and cascaded selection (Chen et al., 2024). Yet, a growing body of evidence suggests that naively adding more models does not guarantee performance improvement: recent scaling studies (Kim et al., 2025) observe diminishing or even negative returns from coordination once single-agent accuracy exceeds $\sim$45%, while systematic failure analysis (Cemri et al., 2025) identifies inter-agent misalignment as a dominant failure mode in multi-agent systems. These observations point to error correlation as the culprit. To address this, a recent study (Ai et al., 2025) leverages second-order statistics for optimal weight aggregation, LLM-TOPLA (Tekin et al., 2024) introduces focal diversity metrics for ensemble pruning, and MUSE (Kruse et al., 2025) applies Jensen-Shannon divergence to select well-calibrated LLM subsets. Additionally, (Turkmen et al., 2025) study the accuracy-timeliness tradeoffs in networked LLM systems under adaptive majority voting. In this work, we take a fundamentally different approach: rather than designing better aggregation rules, we ask which models to select in the first place, and provide a principled probabilistic framework based on Gaussian-copula correlation structure.

Information-Theoretic Feature Selection: The idea of greedily selecting elements to maximize mutual information originates in feature selection, where the seminal mRMR criterion (Peng et al., 2005) balances relevance against redundancy. Brown et al. (Brown et al., 2012) later unified a decade of variants under a common framework: all approximate conditional mutual information (CMI) and all penalize inter-feature correlation as redundancy. Recent work extends this to dynamic settings via learned CMI policies (Covert et al., 2023; Gadgil et al., 2024), yet the core assumption persists. We observe that this intuition does not transfer to ensemble selection.

Gaussian-Copula Models: Gaussian-copulas provide a flexible framework for modeling dependencies while preserving arbitrary marginals, and have found widespread use in various applications from credit risk modeling (Li, 2000) to mixed data analysis (Nelsen, 2007). While recent work (Pan et al., 2025) applies bivariate Gaussian-copula to model refusal behavior within a single LLM, another work, (Elumar et al., 2025), uses a Gaussian-copula sampler to model dependent correctness for Monte Carlo confidence estimation. In this work, we model cross-LLM ensemble errors with a Gaussian-copula and leverage it for an information-theoretic saturation theorem and mutual information (MI)-driven greedy selection.

B. Proof of Theorem 3.1

We first introduce the necessary definitions and then proceed with the proof.

Definition B.1 (M-ary Symmetric Channel)**.** A M-ary symmetric channel (MSC) with error rate $\epsilon \in [0, (M-1)/M)$ is a memoryless channel with input $Y \in [M]$ and output $X \in [M]$ specified by the transition kernel:

$$\mathbb{P}(X = y' \mid Y = y) = \begin{cases} 1-\epsilon, & y' = y, \\ \dfrac{\epsilon}{M-1}, & y' \neq y. \end{cases} \tag{16}$$

Equivalently, with probability $1-\epsilon$ the output equals the input, and with probability ϵ it is drawn uniformly at random from the other $M-1$ symbols. We express the overall error probability as $\mathbb{P}(X \neq Y) = \epsilon$.

Definition B.2 (Stochastic Degradation)**.** A (communication) channel $\mathcal{C}'$ is said to be a *stochastically degraded* version of channel $\mathcal{C}$ if there exists an intermediate channel $\mathcal{C}''$ such that the cascade $\mathcal{C} \to \mathcal{C}''$ is statistically equivalent to $\mathcal{C}'$. In other words, the output of $\mathcal{C}'$ can be obtained by further corrupting the output of $\mathcal{C}$ through an independent noise process.

Lemma B.3 (MSC Degradation)**.** *Let $\mathcal{C}_1$ be a MSC with error rate ϵ_1 and $\mathcal{C}_2$ be a MSC with error rate ϵ_2 where $0 \leq \epsilon_1 \leq \epsilon_2 < (M-1)/M$. Then, $\mathcal{C}_2$ is a stochastically degraded version of $\mathcal{C}_1$. Specifically, $\mathcal{C}_2$ can be constructed by cascading $\mathcal{C}_1$ with an independent MSC having an error rate:*

$$\delta = \frac{\epsilon_2 - \epsilon_1}{1 - \dfrac{M\,\epsilon_1}{M-1}}. \tag{17}$$

Proof. Let $Y \in [M]$ be the input, Z_1 be the output of $\mathcal{C}_1$, and Z_2 be the output after passing Z_1 through a MSC with error rate δ. By the symmetry of both channels, the end-to-end transition kernel is itself symmetric, so it suffices to verify the diagonal transition probability $\mathbb{P}(Z_2 = y \mid Y = y)$.

Conditioning on Z_1 and using the independence of the two channel uses, we write the probability for the cascaded channel as:

$$\begin{aligned} \mathbb{P}(Z_2 = y \mid Y = y) &= \mathbb{P}(Z_2 = y \mid Z_1 = y)\mathbb{P}(Z_1 = y \mid Y = y) \\ &\quad + \sum_{y'' \neq y} \mathbb{P}(Z_2 = y \mid Z_1 = y'')\mathbb{P}(Z_1 = y'' \mid Y = y) \\ &= (1-\delta)(1-\epsilon_1) + (M-1) \cdot \frac{\delta}{M-1} \cdot \frac{\epsilon_1}{M-1} \\ &= (1-\delta)(1-\epsilon_1) + \frac{\delta\,\epsilon_1}{M-1}. \end{aligned} \tag{18}$$

Setting this equal to $1-\epsilon_2$ and solving for δ yields:

$$\begin{aligned} 1 - \epsilon_1 - \delta + \delta\,\epsilon_1 + \frac{\delta\,\epsilon_1}{M-1} &= 1 - \epsilon_2 \\ -\epsilon_1 - \delta\left[1 - \epsilon_1\left(1 + \tfrac{1}{M-1}\right)\right] &= -\epsilon_2 \\ \delta\left[1 - \frac{M\,\epsilon_1}{M-1}\right] &= \epsilon_2 - \epsilon_1. \end{aligned} \tag{19}$$

This directly yields (17). Since $\epsilon_1 < (M-1)/M$, the denominator $1 - M\epsilon_1/(M-1)$ is strictly positive, and since $\epsilon_1 \leq \epsilon_2$, the numerator is non-negative, confirming $\delta \geq 0$. Factoring the bounds ensures $\delta < (M-1)/M$, confirming the validity of this construction. □

Definition B.4 (Top-k Ensemble)**.** Given m channels indexed by increasing error rates $\epsilon_1 \leq \epsilon_2 \leq \cdots \leq \epsilon_m$, the *top-$k$ ensemble* is defined as $H_k = \{1, 2, \ldots, k\}$, containing the k channels with the smallest error rates.

We now restate and prove the main theorem.

Theorem 3.1 [Alignment of Error and Mutual Information] Consider an ensemble of m independent MSCs where channel j has error rate $\epsilon_j < (M-1)/M$, indexed such that $\epsilon_1 \leq \epsilon_2 \leq \cdots \leq \epsilon_m$, and an arbitrary class prior $\boldsymbol{\pi}$ with full support on $[M]$. For any subset $S \subseteq [m]$ with $|S| = k$, the top-k set H_k satisfies:

$$I(Y; \mathbf{X}_{\mathbf{H_k}}) \geq I(Y; \mathbf{X}_{\mathbf{S}}), \tag{20}$$

$$\mathbb{P}_e(H_k) \leq \mathbb{P}_e(S). \tag{21}$$

Proof. We prove this result by constructing an explicit coupling between $\mathbf{X}_{\mathbf{H_k}}$ and $\mathbf{X}_{\mathbf{S}}$, demonstrating that $\mathbf{X}_{\mathbf{S}}$ can be viewed as a degraded version of $\mathbf{X}_{\mathbf{H_k}}$.

We now define a bijection $\pi : S \to H_k$ satisfying $\epsilon_{\pi(j)} \leq \epsilon_j$. To see this, first set $\pi(j) = j$ for all $j \in S \cap H_k$. The remaining elements satisfy $|S \setminus H_k| = |H_k \setminus S|$, so we can bijectively map $S \setminus H_k$ onto $H_k \setminus S$. For any $j \in S \setminus H_k$ we have $j > k$, hence $\epsilon_j \geq \epsilon_k \geq \epsilon_i$ for every $i \in H_k$, so in particular $\epsilon_{\pi(j)} \leq \epsilon_j$.

Using the bijection π from the step above, we now construct a random vector $\widetilde{X}_S = (\widetilde{X}_j)_{j \in S}$ as follows: for each $j \in S$, let $i = \pi(j) \in H_k$ be the paired index with $\epsilon_i \leq \epsilon_j$, compute the degradation parameter $\delta_j = \frac{\epsilon_j - \epsilon_i}{1 - M\epsilon_i/(M-1)}$ using Lemma B.3, and generate $\widetilde{X}_j$ by passing X_i through an independent MSC with error rate δ_j. From Lemma B.3, each $\widetilde{X}_j$ has the same marginal distribution as X_j, i.e., it is the output of a MSC with error rate ϵ_j. We now argue that $\widetilde{\mathbf{X}}_{\mathbf{S}} \overset{d}{=} \mathbf{X}_{\mathbf{S}}$: each $\widetilde{X}_j$ has the correct marginal channel kernel $\mathrm{MSC}(\epsilon_j)$, the degradation noises are independent across components, and the original channels are conditionally independent given Y. This establishes that the joint distribution of $\widetilde{X}_S$ matches that of $\mathbf{X}_S$, i.e.,

$$\widetilde{\mathbf{X}}_{\mathbf{S}} \overset{d}{=} \mathbf{X}_{\mathbf{S}}. \tag{22}$$

By construction, $\widetilde{X}_S$ is a (possibly randomized) function of X_{H_k}, with randomization independent of Y. As a result, we have the Markov chain:

$$Y \longrightarrow \mathbf{X}_{\mathbf{H_k}} \longrightarrow \widetilde{\mathbf{X}}_{\mathbf{S}} \overset{d}{=} \mathbf{X}_{\mathbf{S}}. \tag{23}$$

Applying the Data Processing Inequality (Cover & Thomas, 2006) to the Markov chain in Equation (23), we obtain:

$$I(Y; \mathbf{X}_{\mathbf{S}}) = I(Y; \widetilde{\mathbf{X}}_{\mathbf{S}}) \leq I(Y; \mathbf{X}_{\mathbf{H_k}}), \tag{24}$$

which proves (20).

Now, for the error probability, any decision rule $\hat{Y} : [M]^{|S|} \to [M]$ based on $\mathbf{X}_{\mathbf{S}}$ can be replicated using $\mathbf{X}_{\mathbf{H_k}}$ by defining a composed rule $\hat{Y}'(\mathbf{X}_{\mathbf{H_k}}) \triangleq \hat{Y}(\widetilde{\mathbf{X}}_{\mathbf{S}})$, where $\widetilde{\mathbf{X}}_{\mathbf{S}}$ is generated through the degradation construction. Since $\widetilde{\mathbf{X}}_{\mathbf{S}} \overset{d}{=} \mathbf{X}_{\mathbf{S}}$, the composed rule attains the same Bayes risk:

$$\mathbb{P}_e(S) = \min_{\hat{Y}} \mathbb{P}(\hat{Y}(\mathbf{X}_{\mathbf{S}}) \neq Y) = \min_{\hat{Y}} \mathbb{P}(\hat{Y}(\widetilde{\mathbf{X}}_{\mathbf{S}}) \neq Y). \tag{25}$$

Because the optimal MAP rule on $\mathbf{X}_{\mathbf{H_k}}$ achieves Bayes risk no larger than that of any rule on $\mathbf{X}_{\mathbf{S}}$ (including the MAP rule on $\mathbf{X}_{\mathbf{S}}$), we have:

$$\mathbb{P}_e(H_k) = \min_{\hat{Y}} \mathbb{P}(\hat{Y}(\mathbf{X}_{\mathbf{H_k}}) \neq Y) \leq \mathbb{P}_e(S), \tag{26}$$

which proves (21). Crucially, monotonicity of error probability under stochastic degradation is prior-agnostic, allowing this to hold for any arbitrary class prior $\boldsymbol{\pi}$. □

Consequently, we make the following remarks regarding this result:

Remark B.5 (Tightness of the Bound). The inequalities in Theorem 3.1 become equalities whenever S consists of k channels with the smallest error rates (ties allowed). For any other subset $S \neq H_k$, at least one component must be strictly degraded (i.e., $\delta_j > 0$ for some j), leading to a strict inequality.

Remark B.6 (Necessity of Symmetry). The symmetry of the noise model is essential. For asymmetric M-ary channels, two channels with identical accuracy can carry different mutual information about Y. For example, under $\boldsymbol{\pi} = (0.9, 0.05, 0.05)$, a channel that always confuses class 1 with class 2 is far less informative than one that distributes its output uniformly when wrong, even at identical ϵ_j. Under the symmetric MSC model, however, accuracy is a sufficient statistic for ranking.

Remark B.7 (Recovery of the Binary Case). Setting $M = 2$ in Lemma B.3 gives $\delta = (\epsilon_2 - \epsilon_1)/(1 - 2\epsilon_1)$, which perfectly recovers the classical BSC degradation formula. Setting $M = 2$ and $\boldsymbol{\pi} = (1/2, 1/2)$ recovers the binary balanced statement.

To visualize the stochastic degradation argument, we provide the following example.

Example B.1: Stochastic Degradation Illustration

Consider a pool of $m = 3$ independent models classifying into $M = 3$ classes with error rates: $\epsilon_1 = 0.1$ (expert), $\epsilon_2 = 0.2$ (average), and $\epsilon_3 = 0.4$ (novice).
We seek the optimal subset of size $k = 2$, which is the top-k subset $H_2 = \{1, 2\}$. As a suboptimal comparison, consider $S = \{1, 3\}$.
Bijection Construction. We construct the mapping $\pi : S \to H_2$:

- $\pi(1) = 1$: Model 1 maps to itself ($\epsilon_1 \leq \epsilon_1$).
- $\pi(3) = 2$: Model 3 maps to Model 2 ($\epsilon_2 \leq \epsilon_3$).

Degradation Parameters. For each mapping, we compute the crossover probability using (17):

- $\delta_1 = 0$: No noise injection needed (identical mapping).
- $\delta_3 = \frac{\epsilon_3 - \epsilon_2}{1 - \frac{M\,\epsilon_2}{M-1}} = \frac{0.4-0.2}{1-\frac{3(0.2)}{2}} = \frac{0.2}{1-0.3} = \frac{2}{7} \approx 0.286$: Multiply X_2 by an independent MSC noise process with error rate $\delta \approx 0.286$.

Resulting Markov Chain. This construction establishes:

$$Y \longrightarrow \mathbf{X}_{H_2} \xrightarrow{\text{MSC}(\delta=2/7)} \mathbf{X}_S$$

By the Data Processing Inequality, $I(Y; \mathbf{X}_{H_2}) \geq I(Y; \mathbf{X}_S)$. Similarly, $\mathbb{P}_e(H_2) \leq \mathbb{P}_e(S)$, confirming that the top-k ensemble is optimal.

C. Proof of Theorem 3.3

We first establish the key technical tools used in the proof.

Definition C.1 (The Conditional Mutual Information)**.** The conditional mutual information between random variables X and Y given Z is defined as:

$$I(X;Y \mid Z) \triangleq H(X \mid Z) - H(X \mid Y,Z) = \mathbb{E}_Z \left[D_{\mathrm{KL}}(P_{X,Y|Z} \| P_{X|Z} P_{Y|Z}) \right]. \tag{27}$$

This quantity measures the information that Y provides about X beyond what is already known from Z.

Lemma C.2 (The Chain Rule for Mutual Information)**.** *For any random variables X, Y, and Z, we have:*

$$I(X;Y,Z) = I(X;Z) + I(X;Y \mid Z) = I(X;Y) + I(X;Z \mid Y). \tag{28}$$

Lemma C.3 (Entropy of M-ary Errors)**.** *Let $Y \in [M]$ be the true label, $X \in [M]$ be the prediction from an M-ary Symmetric Channel (MSC), and $E = \mathbb{1}(X \neq Y)$ be the error indicator with $\mathbb{P}(E=1) = \epsilon$. The conditional entropy of the prediction given the true label decomposes as:*

$$H(X \mid Y) = H(E \mid Y) + \epsilon \log(M-1). \tag{29}$$

Proof. By the chain rule for entropy, $H(X \mid Y) = H(X, E \mid Y)$ because E is a deterministic function of X and Y. We can expand this as $H(E \mid Y) + H(X \mid E, Y)$. To compute $H(X \mid E, Y)$, we condition on the value of E:

- If $E = 0$, $X = Y$ exactly, leaving no uncertainty: $H(X \mid E = 0, Y) = 0$.
- If $E = 1$, the MSC assumption dictates that X is uniformly distributed over the remaining $M-1$ classes: $H(X \mid E = 1, Y) = \log(M-1)$.

Taking the expectation over E, we obtain $H(X \mid E, Y) = \mathbb{P}(E=1)\log(M-1) = \epsilon \log(M-1)$. Adding this to $H(E \mid Y)$ yields the result. □

For convenience, we first restate the decomposition theorem (Theorem 3.3) and then provide its proof.

Theorem 3.3 [Accuracy-Redundancy-Error Decomposition] For outputs $X_j \in [M]$ and true label $Y \in [M]$ drawn from an arbitrary prior $\boldsymbol{\pi}$, let $E_j = \mathbb{1}(X_j \neq Y)$ be the error indicator for model j, and define the marginal information gain of adding model j to subset S as:

$$\Delta(j \mid S) \triangleq I(Y; X_j \mid \mathbf{X_S}). \tag{30}$$

Then, $\Delta(j \mid S)$ admits the decomposition:

$$\Delta(j \mid S) = \underbrace{I(Y;X_j)}_{\text{Accuracy}} - \underbrace{I(X_j;\mathbf{X_S})}_{\text{Redundancy}} + \underbrace{I(E_j;E_S)}_{\text{Error Correlation}} + \underbrace{\Lambda_j(S)}_{\text{Correction}}, \tag{31}$$

where the label-dependent correction term is given by:

$$\Lambda_j(S) \triangleq I(E_j; Y \mid E_S) - I(E_j; Y). \tag{32}$$

Proof. We proceed through a series of algebraic manipulations using the chain rule and the M-ary entropy decomposition.

By using the definition of the conditional mutual information, we have:

$$\Delta(j \mid S) = I(Y; X_j \mid \mathbf{X_S}) = H(X_j \mid \mathbf{X_S}) - H(X_j \mid \mathbf{X_S}, Y). \tag{33}$$

Applying the identity $H(X \mid Z) = H(X) - I(X;Z)$, we have:

$$H(X_j \mid \mathbf{X_S}) = H(X_j) - I(X_j; \mathbf{X_S}). \tag{34}$$

Next, we evaluate the term $H(X_j \mid \mathbf{X_S}, Y)$. In our Gaussian-copula framework, the correlation among models is fully captured by the latent error indicators E_S. If a model makes an error ($E_j = 1$), its specific choice among the $M-1$ wrong

classes is an independent uniform draw. Because knowing $\mathbf{X_S}$ and Y perfectly determines E_S, applying the logic from Lemma C.3 to the conditioned variables yields:

$$H(X_j \mid \mathbf{X_S}, Y) = H(E_j \mid E_S, Y) + \epsilon_j \log(M-1). \tag{35}$$

Using the chain rule for entropy on the indicator term, we obtain:

$$H(E_j \mid E_S, Y) = H(E_j \mid Y) - I(E_j; E_S \mid Y). \tag{36}$$

Substituting (34), (35), and (36) into (33), we have:

$$\Delta(j \mid S) = H(X_j) - I(X_j; \mathbf{X_S}) - H(E_j \mid Y) + I(E_j; E_S \mid Y) - \epsilon_j \log(M-1). \tag{37}$$

From Lemma C.3, we also know that $H(X_j \mid Y) = H(E_j \mid Y) + \epsilon_j \log(M-1)$. We can relate this to the marginal accuracy $I(Y; X_j)$ as follows:

$$I(Y; X_j) = H(X_j) - H(X_j \mid Y) = H(X_j) - \left[H(E_j \mid Y) + \epsilon_j \log(M-1)\right]. \tag{38}$$

Rearranging to solve for $H(X_j)$:

$$H(X_j) = I(Y; X_j) + H(E_j \mid Y) + \epsilon_j \log(M-1). \tag{39}$$

Substituting Equation (39) into Equation (37), observe that the error dispersion terms $\epsilon_j \log(M-1)$ perfectly cancel out:

$$\begin{aligned}\Delta(j \mid S) &= \Big[I(Y; X_j) + H(E_j \mid Y) + \epsilon_j \log(M-1)\Big] \\ &\quad - I(X_j; \mathbf{X_S}) - H(E_j \mid Y) + I(E_j; E_S \mid Y) - \epsilon_j \log(M-1) \\ &= I(Y; X_j) - I(X_j; \mathbf{X_S}) + I(E_j; E_S \mid Y).\end{aligned} \tag{40}$$

Applying Lemma C.2 to $I(E_j; E_S, Y)$ in two ways:

$$I(E_j; E_S, Y) = I(E_j; E_S) + I(E_j; Y \mid E_S), \tag{41}$$

$$I(E_j; E_S, Y) = I(E_j; Y) + I(E_j; E_S \mid Y). \tag{42}$$

Equating Equation (41) and Equation (42) and solving for $I(E_j; E_S \mid Y)$:

$$I(E_j; E_S \mid Y) = I(E_j; E_S) + \underbrace{\left[I(E_j; Y \mid E_S) - I(E_j; Y)\right]}_{\Lambda_j(S)}. \tag{43}$$

Substituting Equation (43) into the canceled result from Equation (40):

$$\Delta(j \mid S) = I(Y; X_j) - I(X_j; \mathbf{X_S}) + I(E_j; E_S) + \Lambda_j(S). \tag{44}$$

This completes the proof. □

Remark C.4 (Interpretation of Terms). The decomposition in Equation (31) has an intuitive interpretation:

- **Accuracy** $I(Y; X_j)$: The single predictive power of model j.
- **Redundancy** $I(X_j; \mathbf{X_S})$: Information overlap between model j and the existing ensemble S.
- **Error Correlation** $I(E_j; E_S)$: dependence among model errors, capturing shared failure modes that influences the incremental value of adding model j.
- **Correction** $\Lambda_j(S)$: Adjustment for label-dependent error behavior.

Corollary C.5 (Label-Invariant Case)**.** *Under the label-invariant error assumption* $(E_1, \ldots, E_m) \perp Y$*, we have:*

$$I(E_j; Y) = 0 \quad \textit{and} \quad I(E_j; Y \mid E_S) = 0, \quad \forall j, S. \tag{45}$$

Consequently, $\Lambda_j(S) = 0$ *for all* j *and* S*, and the decomposition simplifies to:*

$$\Delta(j \mid S) = I(Y; X_j) - I(X_j; \mathbf{X_S}) + I(E_j; E_S). \tag{46}$$

Proof. Independence $(E_1, \ldots, E_m) \perp Y$ implies $I(E_j; Y) = 0$ directly. For the conditional term, note that E_S is a function of $(E_1, \ldots, E_m)$, so $(E_j, E_S) \perp Y$, which implies $I(E_j; Y \mid E_S) = 0$. □

Proposition C.6 (Bound on the Correction Term)**.** *The label-dependence correction term satisfies:*

$$-I(E_j; Y) \leq \Lambda_j(S) \leq H(Y) - I(E_j; Y). \tag{47}$$

Proof. By non-negativity of mutual information, $I(E_j; Y \mid E_S) \geq 0$. Therefore:

$$\Lambda_j(S) = I(E_j; Y \mid E_S) - I(E_j; Y) \geq -I(E_j; Y). \tag{48}$$

Conditional mutual information is bounded by conditional entropy:

$$I(E_j; Y \mid E_S) \leq H(Y \mid E_S) \leq H(Y). \tag{49}$$

Therefore:

$$\Lambda_j(S) = I(E_j; Y \mid E_S) - I(E_j; Y) \leq H(Y) - I(E_j; Y). \tag{50}$$

□

Remark C.7 (Practical Implications). For an M-ary classification task, $H(Y) \leq \log_2(M)$ bits. Moreover, if the *joint* label-dependence is small, e.g.,

$$I(E_j, E_S; Y) \leq \delta, \tag{51}$$

then $|\Lambda_j(S)| \leq \delta$ and the correction vanishes as $\delta \to 0$. This motivates ignoring $\Lambda_j(S)$ when empirical estimates indicate weak joint dependence between error patterns and labels.

D. Proof of Theorem 3.4

First, we restate the Gaussian-copula (one-factor) representation used in Section 2.1 adapted to the M-ary symmetric channel model, and then give a step-by-step proof.

D.1. Model recap and notation

Let $Y \in [M]$ be the M-ary label with a balanced prior $\mathbb{P}(Y = y) = 1/M$ for all $y \in [M]$. Let $U \sim \mathcal{N}(0, 1)$ be a latent common factor and let $V_1, V_2, \dots$ be i.i.d. $\mathcal{N}(0, 1)$, independent of U and Y. Fix $\rho \in (0, 1)$ and define latent Gaussian scores

$$T_j \triangleq \sqrt{\rho}\, U + \sqrt{1-\rho}\, V_j, \qquad j = 1, 2, \dots \tag{52}$$

so that $(T_1, \dots, T_K)$ is equicorrelated Gaussian with $\mathrm{corr}(T_i, T_j) = \rho$ for $i \neq j$.

Define the error indicator by thresholding:

$$E_j \triangleq \mathbb{1}(T_j \leq \tau). \tag{53}$$

Conditional on $Y = y$ and $E_j = 0$, the observed output is $X_j = y$; conditional on $Y = y$ and $E_j = 1$, X_j is drawn uniformly from $[M] \setminus \{y\}$, with the wrong-label draws conditionally independent across j given $(E_1, \dots, E_K)$ and Y. This is the M-ary symmetric channel coupling (1) introduced in Section 2.1.

The threshold τ is chosen to match the marginal accuracy $\alpha > 1/M$: since $T_j \sim \mathcal{N}(0, 1)$ marginally, we have

$$\mathbb{P}(E_j = 1) = \mathbb{P}(T_j \leq \tau) = \Phi(\tau) = 1 - \alpha. \tag{54}$$

Then, we can obtain $\tau = \Phi^{-1}(1 - \alpha)$.

D.2. Proof of Theorem 3.4

Conditioned on $U = u$, we have from (52) that

$$T_j \mid (U = u) \;\sim\; \mathcal{N}\big(\sqrt{\rho}\, u,\, 1 - \rho\big) \quad \text{independently across } j.$$

Therefore, conditioned on $U = u$, the errors $(E_j)_{j \geq 1}$ are i.i.d. Bernoulli with

$$p(u) \triangleq \mathbb{P}(E_j = 1 \mid U = u) = \mathbb{P}(T_j \leq \tau \mid U = u) = \Phi\left(\frac{\tau - \sqrt{\rho}\, u}{\sqrt{1-\rho}}\right). \tag{55}$$

For each class $y \in [M]$, let $\bar{N}_K(y) \triangleq \frac{1}{K} \sum_{j=1}^{K} \mathbb{1}(X_j = y)$ denote the empirical fraction of votes for class y. Conditional on $U = u$ and $Y = y^\star$ (the true label), each model votes for $y^\star$ with probability $1 - p(u)$ and for each $y \neq y^\star$ with probability $p(u)/(M-1)$, conditionally independently across j by the M-ary symmetric channel coupling. By the Strong Law of Large Numbers applied conditional on (U, Y), we have almost surely

$$\bar{N}_K(y^\star) \xrightarrow[K\to\infty]{a.s.} 1 - p(U), \qquad \bar{N}_K(y) \xrightarrow[K\to\infty]{a.s.} \frac{p(U)}{M-1} \text{ for each } y \neq y^\star. \tag{56}$$

Under the balanced prior, the MAP decision rule reduces to plurality voting: $\widehat{Y}_{\mathrm{MAP}} = \arg\max_{y \in [M]} \bar{N}_K(y)$, with ties broken arbitrarily. From (56), plurality recovers $y^\star$ asymptotically if and only if the empirical fraction for $y^\star$ exceeds that of every other class, which in the limit is the condition

$$1 - p(U) > \frac{p(U)}{M-1} \quad \Longleftrightarrow \quad p(U) < \frac{M-1}{M}. \tag{57}$$

Since the asymptotic failure event depends only on $p(U)$, we obtain the almost sure limit:

$$\mathbb{1}\{\widehat{Y}_{\mathrm{MAP}} \neq Y\} \xrightarrow[K\to\infty]{a.s.} \mathbb{1}\left\{p(U) > \frac{M-1}{M}\right\}. \tag{58}$$

Taking expectations yields

$$\lim_{K\to\infty} \mathbb{P}(\widehat{Y}_{\mathrm{MAP}} \neq Y) = \mathbb{P}\left(p(U) > \frac{M-1}{M}\right). \tag{59}$$

Now let us find the probability of the event $p(U) > (M-1)/M$ explicitly. From (55), using that $\Phi(\cdot)$ is strictly increasing, $p(U) > (M-1)/M$ is equivalent to having

$$\frac{\tau - \sqrt{\rho}\,U}{\sqrt{1-\rho}} > \Phi^{-1}\left(\frac{M-1}{M}\right).$$

This condition implies

$$U < \frac{\tau - \sqrt{1-\rho}\,\Phi^{-1}((M-1)/M)}{\sqrt{\rho}}.$$

Since $U \sim \mathcal{N}(0,1)$, we obtain the following

$$\mathbb{P}\left(p(U) > \frac{M-1}{M}\right) = \Phi\left(\frac{\tau - \sqrt{1-\rho}\,\Phi^{-1}((M-1)/M)}{\sqrt{\rho}}\right). \tag{60}$$

Substituting $\tau = \Phi^{-1}(1-\alpha)$ using (54) in (60) gives

$$\lim_{K\to\infty} \mathbb{P}(\widehat{Y}_{\mathrm{MAP}} \neq Y) = \Phi\left(\frac{\Phi^{-1}(1-\alpha) - \sqrt{1-\rho}\,\Phi^{-1}((M-1)/M)}{\sqrt{\rho}}\right),$$

which is strictly positive for any $\rho > 0$ and $\alpha \in (1/M, 1)$. This completes the proof.

Remark D.1 (Edge cases and interpretation). Because $\Phi^{-1}(1-\alpha) < \Phi^{-1}((M-1)/M)$ for $\alpha > 1/M$, the numerator inside $\Phi(\cdot)$ is dominated by a negative term, and the floor lies in $(0, (M-1)/M)$ for any $\rho > 0$. As ρ decreases to 0, the argument tends to $-\infty$ and $\lim_{K\to\infty} \mathbb{P}(\widehat{Y}_{\mathrm{MAP}} \neq Y)$ vanishes, recovering the independent-error regime. As ρ increases to 1, the argument approaches $\Phi^{-1}(1-\alpha)$, which implies $\lim_{K\to\infty} \mathbb{P}(\widehat{Y}_{\mathrm{MAP}} \neq Y) = 1-\alpha$, matching the intuition that perfectly correlated models provide no ensemble gain. Setting $M = 2$ gives $\Phi^{-1}(1/2) = 0$, and the bound reduces to the binary closed form $\Phi(\Phi^{-1}(1-\alpha)/\sqrt{\rho})$ used in our experiments.

E. Proof of Corollary 3.5

We now extend the proof of Theorem 3.4 to an arbitrary class prior $\boldsymbol{\pi}$ on $[M]$ with full support, i.e., $\pi_y > 0$ for all $y \in [M]$.

E.1. Proof of Theorem 3.5

Under prior $\boldsymbol{\pi}$, the MAP decision rule selects

$$\widehat{Y}_{\mathrm{MAP}} = \arg\max_{y\in[M]} \{\log \pi_y + \ell_K(y)\}, \tag{61}$$

where $\ell_K(y) \triangleq \log \mathbb{P}(\mathbf{X}_{[\mathbf{K}]} \mid Y = y)$ is the conditional log-likelihood of the observed ensemble outputs under hypothesis $Y = y$. Conditional on $(U, Y) = (u, y^\star)$, the empirical fractions $\bar{N}_K(y)$ converge as in (56), and by the conditional independence of the M-ary symmetric channel coupling we can write

$$\ell_K(y) = \sum_{j=1}^{K} \mathbb{1}(X_j = y) \log(1-\epsilon) + \sum_{j=1}^{K} \mathbb{1}(X_j \neq y) \log \frac{\epsilon}{M-1}, \tag{62}$$

where $\epsilon = 1 - \alpha$. Substituting the empirical fractions and using (56), the conditional log-likelihood per model converges almost surely:

Substituting the empirical fractions and using (56), the conditional log-likelihood per model converges almost surely:

$$\frac{1}{K}\ell_K(y) \xrightarrow[K\to\infty]{a.s.} L(y; U, y^\star), \tag{63}$$

where the limit $L(y; U, y^\star)$ takes the explicit form

$$L(y; U, y^\star) = \begin{cases} (1 - p(U))\, \log(1-\epsilon) + p(U)\, \log \dfrac{\epsilon}{M-1}, & y = y^\star, \\ \dfrac{p(U)}{M-1}\, \log(1-\epsilon) + \left(1 - \dfrac{p(U)}{M-1}\right) \log \dfrac{\epsilon}{M-1}, & y \neq y^\star. \end{cases} \tag{64}$$

The first case follows because, conditional on $(U, Y) = (u, y^\star)$, the empirical fraction of correct outputs $\bar{N}_K(y^\star)$ converges to $1 - p(U)$ and the empirical fraction of incorrect outputs converges to $p(U)$. The second case follows similarly: under hypothesis $Y = y \neq y^\star$, an output $X_j = y$ is treated as correct (probability mass given by the empirical fraction $\bar{N}_K(y) \to p(U)/(M-1)$), and all other outputs are treated as incorrect (mass $1 - p(U)/(M-1)$).

Subtracting the two cases of (64), the asymptotic per-sample log-likelihood gap between the truth and any alternative is

$$L(y^\star; U, y^\star) - L(y; U, y^\star) = \left(1 - p(U) - \frac{p(U)}{M-1}\right) \log \frac{(1-\epsilon)(M-1)}{\epsilon}, \qquad y \neq y^\star. \tag{65}$$

The first factor is $1 - p(U)\, M/(M-1)$, which is strictly positive whenever $p(U) < (M-1)/M$ and zero on the boundary; the second factor is $\log((1-\epsilon)(M-1)/\epsilon) > 0$ since $\epsilon = 1 - \alpha < (M-1)/M$ implies $(1-\epsilon)(M-1)/\epsilon > 1$. Hence $L(y^\star; U, y^\star) > L(y; U, y^\star)$ for every $y \neq y^\star$ in the regime $p(U) < (M-1)/M$, with the gap shrinking to zero as $p(U) \to (M-1)/M$.where $L(y; U, y^\star)$ is a deterministic function of $(U, y^\star)$ that takes its unique maximum at $y = y^\star$ whenever $p(U) < (M-1)/M$ (the "decisively correct" regime), with $L(y^\star; U, y^\star) - L(y; U, y^\star)$ strictly positive for all $y \neq y^\star$ in this regime.

Returning to the MAP rule (61), the difference between the score for $y^\star$ and any alternative y is

$$\big[\log \pi_{y^\star} + \ell_K(y^\star)\big] - \big[\log \pi_y + \ell_K(y)\big] = \underbrace{\log \frac{\pi_{y^\star}}{\pi_y}}_{\text{finite offset}} + \underbrace{\ell_K(y^\star) - \ell_K(y)}_{=K\cdot[L(y^\star;U,y^\star)-L(y;U,y^\star)]+o(K)}. \tag{66}$$

The first term on the right-hand side is a finite constant (since $\pi_y > 0$ for all y), while the second term grows linearly in K at a strictly positive rate whenever $p(U) < (M-1)/M$. Hence, in the regime $p(U) < (M-1)/M$, the score for $y^\star$ exceeds the score for every alternative y for all K sufficiently large, almost surely; conversely, in the regime $p(U) > (M-1)/M$,

the empirical fractions (56) place mass below $(M-1)/M$ on $y^\star$ and above $1/M$ on multiple alternative classes, so the MAP rule fails to recover $y^\star$.

Combining the two regimes, we obtain the almost sure limit:

$$\mathbb{1}\{\widehat{Y}_{\mathrm{MAP}} \neq Y\} \xrightarrow[K\to\infty]{a.s.} \mathbb{1}\left\{p(U) > \frac{M-1}{M}\right\}, \tag{67}$$

which is identical to (58) in the balanced case. Taking expectations and applying the same calculation as in (60) yields

$$\lim_{K\to\infty} \mathbb{P}(\widehat{Y}_{\mathrm{MAP}} \neq Y) = \Phi\left(\frac{\Phi^{-1}(1-\alpha) - \sqrt{1-\rho}\,\Phi^{-1}((M-1)/M)}{\sqrt{\rho}}\right),$$

which is the same expression as in Theorem 3.4. This completes the proof.

Remark E.1 (Rate of convergence). While the asymptotic floor in Theorem 3.5 is identical to the balanced case, the prior $\boldsymbol{\pi}$ does affect the *rate* at which the limit is approached. Specifically, the finite log-prior offset $\log(\pi_{y^\star}/\pi_y)$ in (66) can take either sign, so for unfavorable realizations of $(U, y^\star)$ where $y^\star$ has small prior mass, the linear-in-K likelihood signal must overcome a larger initial offset before the MAP rule recovers $y^\star$. The rate of convergence is governed by $\min_{y\neq y^\star}[L(y^\star; U, y^\star) - L(y; U, y^\star)]$ relative to the prior offset, and is slowest when $\boldsymbol{\pi}$ is most skewed.

F. Difficulty-Aware Decomposition and the Price of Difficulty

To justify our greedy selection and to further explain Theorem 3.4, we expose a difficulty-aware submodular component and an intrinsic saturation penalty. The decomposition below holds for arbitrary discrete label alphabets and class priors, so it applies directly in both the binary balanced setting used in our experiments and the general M-ary imbalanced setting developed in Section 2.

Theorem F.1 (Difficulty-Aware Decomposition and Price of Difficulty)**.** *Let $Y \in [M]$ be the true label drawn from an arbitrary class prior $\boldsymbol{\pi}$, and let $\mathbf{X}_{[\mathbf{K}]} = \{X_1, \ldots, X_K\}$ be discrete model outputs. Assume that there exists a latent variable D (representing the prompt difficulty) such that:*

1. *$D \perp Y$, i.e., D and Y are statistically independent random variables,*
2. *for every $S \subseteq [K]$, model outputs are conditionally independent given (Y, D):*

$$\mathbb{P}(\mathbf{X_S} \mid Y, D) = \prod_{j \in S} \mathbb{P}(X_j \mid Y, D). \tag{68}$$

Then, for every $S \subseteq [K]$,

$$I(Y; \mathbf{X_S}) = I(Y; \mathbf{X_S} \mid D) - I(Y; D \mid \mathbf{X_S}). \tag{69}$$

We refer to $I(Y; \mathbf{X_S} \mid D)$ as the difficulty-aware mutual information which quantifies how informative the subset would be if the latent difficulty D were revealed to the decoder. We refer to $I(Y; D \mid \mathbf{X_S})$ as the Price of Difficulty which measures the remaining ambiguity that is related to not observing D after seeing the ensemble.

Moreover, $F(S) \triangleq I(Y; \mathbf{X_S} \mid D)$ is monotone submodular in S (under a cardinality constraint $|S| \leq k$). Finally, the Price of Difficulty admits the form

$$I(Y; D \mid \mathbf{X_S}) = H(Y \mid \mathbf{X_S}) - H(Y \mid \mathbf{X_S}, D), \tag{70}$$

i.e., the additional label uncertainty induced by not observing D after seeing the ensemble. The decomposition and the submodularity hold for any discrete label alphabet $[M]$ and any prior $\boldsymbol{\pi}$ with full support, recovering the binary balanced case used in our experiments as $M = 2$ with $\pi_{+1} = \pi_{-1} = 1/2$.

Remark F.2. Theorem F.1 isolates a difficulty-aware component $I(Y; \mathbf{X_S} \mid D)$ that is monotone submodular, hence greedy achieves a $(1 - 1/e)$ approximation guarantee for maximizing this term under $|S| \leq k$ (Krause & Golovin, 2014). In practice, D is unobserved, and the difference between the difficulty-aware information and the true information is exactly the Price of Difficulty $I(Y; D \mid \mathbf{X_S})$. If this term does not vanish as $|S|$ grows, then adding more models cannot fully remove uncertainty about Y, which provides an information-theoretic explanation for the saturation phenomenon characterized in Theorem 3.4. Since the decomposition is alphabet- and prior-agnostic, this saturation explanation holds in the same form across the binary balanced and M-ary imbalanced settings.

F.1. Proof of Theorem F.1

For the proof, we use the conditional mutual information and chain rule as stated in: Definition C.1 and Lemma C.2. The argument uses only properties of mutual information and Shannon entropy that hold for arbitrary discrete random variables, so it applies for any label alphabet $[M]$ and any prior $\boldsymbol{\pi}$. Now, lets proceed with the proof.

Proof. We prove the decomposition, the operational form of the Price of Difficulty, and the monotone submodularity of $F(S) \triangleq I(Y; \mathbf{X_S} \mid D)$.

Let's first derive the decomposition in Equation (69). We can apply the chain rule for mutual information to the triple $(Y, \mathbf{X_S}, D)$ in two different ways:

$$I(Y; \mathbf{X_S}, D) = I(Y; \mathbf{X_S}) + I(Y; D \mid \mathbf{X_S}), \tag{71}$$

$$I(Y; \mathbf{X_S}, D) = I(Y; D) + I(Y; \mathbf{X_S} \mid D). \tag{72}$$

Equating Equations (71) and (72) yields

$$I(Y; \mathbf{X_S}) = I(Y; \mathbf{X_S} \mid D) + I(Y; D) - I(Y; D \mid \mathbf{X_S}). \tag{73}$$

By the assumption in the theorem, $D \perp Y$, hence $I(Y;D)=0$ regardless of the alphabet of Y or the prior $\boldsymbol{\pi}$. Substituting into Equation (73) gives $I(Y;\mathbf{X_S})=I(Y;\mathbf{X_S} \mid D) - I(Y;D \mid \mathbf{X_S})$, which is exactly Equation (69).

Now, we can derive Equation (70) by using the definition of conditional mutual information,

$$I(Y;D \mid \mathbf{X_S})=H(Y \mid \mathbf{X_S}) - H(Y \mid \mathbf{X_S}, D), \tag{74}$$

which matches Equation (70).

To prove the monotone submodularity of $F(S)=I(Y;\mathbf{X_S} \mid D)$, we use an entropy representation:

$$F(S)=I(Y;\mathbf{X_S} \mid D)=H(\mathbf{X_S} \mid D) - H(\mathbf{X_S} \mid Y, D). \tag{75}$$

Under the second assumption, the outputs factorize given (Y,D): $\mathbb{P}(\mathbf{X_S} \mid Y,D)=\prod_{j\in S}\mathbb{P}(X_j \mid Y,D)$. Consequently, the conditional entropy decomposes additively as

$$H(\mathbf{X_S} \mid Y,D)=\sum_{j\in S} H(X_j \mid Y,D), \tag{76}$$

so the set function $S \mapsto H(\mathbf{X_S} \mid Y,D)$ is modular under the conditional independence given (Y,D). Next, for discrete random variables, Shannon entropy is submodular in its set argument, and conditioning preserves submodularity; therefore $S \mapsto H(\mathbf{X_S} \mid D)$ is submodular (see, e.g., (Cover & Thomas, 2006)). These submodularity and modularity properties hold for any discrete alphabet, in particular $[M]$ for any $M \geq 2$. Recalling the identity $F(S)$ in (75), we conclude that F is submodular as a submodular function minus a modular one. Finally, for any $j \notin S$,

$$F(S \cup \{j\}) - F(S)=I(Y;X_j \mid \mathbf{X_S}, D) \geq 0, \tag{77}$$

where the last inequality follows from non-negativity of conditional mutual information. Therefore, F is monotone.

□

G. Implementation Details

In this section, we detail the implementation of the algorithms used throughout the experiments. We first present the smoothed mutual information estimation method, followed by the details of the baseline algorithms. Then, we provide pseudo codes for the MAP, Majority Voting, and Weighted Majority Voting aggregation rules. We conclude with providing the implementation details for the Gaussian-copula parameter estimation that is used for modeling correlated errors.

G.1. Mutual Information Estimation

Algorithm 2 summarizes how we estimate mutual information from finite samples. Given N paired observations of discrete variables A and B, we form empirical marginal and joint counts and convert them to probabilities using Laplace (add-one) smoothing with $\alpha = 1$ (please note that this parameter is for smoothing, not for the accuracy as used in Theorem 3.4). This guarantees strictly positive probability estimates, so the entropy terms are always well-defined (no $\log 0$), and it stabilizes the estimates when some outcomes are rare or absent in the parameter-estimation split (the 80% portion used in each run and split). The mutual information is then computed via the standard identity $I(A;B) = H(A) + H(B) - H(A,B)$, where each entropy term is estimated from the smoothed distributions.

Algorithm 2 Smoothed Mutual Information Estimation

Require: Discrete variables $A \in \mathcal{A}^N$, $B \in \mathcal{B}^N$, smoothing parameter $\alpha > 0$
Ensure: Mutual information estimate $\hat{I}(A;B)$

▷ Compute variable sizes
$K_A \leftarrow |\mathcal{A}|$, $K_B \leftarrow |\mathcal{B}|$, $K_{AB} \leftarrow K_A \cdot K_B$

▷ Estimate marginal distributions with Laplace smoothing
for $a \in \mathcal{A}$ **do**
  $\hat{P}(A = a) \leftarrow \frac{|\{i : A_i = a\}| + \alpha}{N + \alpha K_A}$
end for
for $b \in \mathcal{B}$ **do**
  $\hat{P}(B = b) \leftarrow \frac{|\{i : B_i = b\}| + \alpha}{N + \alpha K_B}$
end for

▷ Estimate joint distribution
for $(a, b) \in \mathcal{A} \times \mathcal{B}$ **do**
  $\hat{P}(A = a, B = b) \leftarrow \frac{|\{i : A_i = a, B_i = b\}| + \alpha}{N + \alpha K_{AB}}$
end for

▷ Compute entropies
$\hat{H}(A) \leftarrow -\sum_{a \in \mathcal{A}} \hat{P}(a) \log_2 \hat{P}(a)$
$\hat{H}(B) \leftarrow -\sum_{b \in \mathcal{B}} \hat{P}(b) \log_2 \hat{P}(b)$
$\hat{H}(A, B) \leftarrow -\sum_{(a,b) \in \mathcal{A} \times \mathcal{B}} \hat{P}(a, b) \log_2 \hat{P}(a, b)$

▷ Mutual information
$\hat{I}(A;B) \leftarrow \max\Big(\hat{H}(A) + \hat{H}(B) - \hat{H}(A,B),\ 0\Big)$

Computational complexity: Let $K_A = |\mathcal{A}|$ and $K_B = |\mathcal{B}|$. Algorithm 2 estimates the marginals in $\mathcal{O}(N + K_A + K_B)$ time (one pass to count plus normalization), estimates the joint in $\mathcal{O}(N + K_A K_B)$ time, and computes entropies in $\mathcal{O}(K_A + K_B + K_A K_B)$ time. Overall, the time complexity is

$$\mathcal{O}\big(N + K_A K_B\big),$$

and the memory cost is $\mathcal{O}(K_A K_B)$ to store the smoothed joint table (plus $\mathcal{O}(K_A + K_B)$ for marginals). In our setting A, B are low-cardinality (e.g., $|\mathcal{A}| = |\mathcal{B}| = 2$ for binary variables), so the MI estimation is negligible compared to LLM query costs.

G.2. Baseline Algorithms

In our experiments, we compare the proposed Greedy MI method against several baseline selection strategies. Based on our implementation, these algorithms operate as follows.

TMC-Shapley (Shapley Value Selection): This method selects models based on their estimated Shapley values, which quantify their average marginal contribution to the ensemble's utility. Because exact computation is exponentially complex, we use a Truncated Monte Carlo (TMC) approximation (Algorithm 3). The algorithm generates a large number of random permutations of the M models. For each permutation, models are added sequentially, and the ensemble's error (evaluated using the MAP aggregator on the training set) is tracked. A model's marginal contribution is the reduction in error it provides when added to the preceding models in the permutation. These contributions are averaged across all permutations to estimate the final Shapley value for each model. The top-k models with the highest Shapley values are selected.

Algorithm 3 TMC-Shapley Selection

Require: Training data $(X^{\text{tr}}, Y^{\text{tr}})$, total models K, desired subset size k, permutations P, tolerance ϵ
Ensure: Selected subset S_{Shapley} of size k
Initialize $v_j \leftarrow 0$ for all $j \in \{1, \ldots, K\}$
$U_{\text{full}} \leftarrow$ MAP_Error$(Y^{\text{tr}}, X^{\text{tr}}, Y^{\text{tr}}, X^{\text{tr}})$
for $p = 1$ to P **do**
  $\pi \leftarrow$ RandomPermutation(M)
  prev_error $\leftarrow 0.5$
  for pos $= 1$ to K **do**
    $j \leftarrow \pi[\text{pos}]$
    if $|\text{prev_error} - U_{\text{full}}| \leq \epsilon$ **then**
      break {Truncation step}
    end if
    $S_{\text{curr}} \leftarrow \pi[1 \ldots \text{pos}]$
    curr_error $\leftarrow$ MAP_Error$(Y^{\text{tr}}, X^{\text{tr}}_{S_{\text{curr}}}, Y^{\text{tr}}, X^{\text{tr}}_{S_{\text{curr}}})$
    $v_j \leftarrow v_j + (\text{prev_error} - \text{curr_error})$
    prev_error $\leftarrow$ curr_error
  end for
end for
$v_j \leftarrow v_j / P$ for all $j \in \{1, \ldots, K\}$
$S_{\text{Shapley}} \leftarrow$ indices of top k models with highest v_j

Logistic Regression Forward Selection (LogReg): This baseline employs a greedy forward-selection strategy paired with a logistic regression classifier (Algorithm 4). Starting with an empty set, the algorithm evaluates all unselected candidate models at each step. For each candidate, it appends the model's predictions to the currently selected subset and fits a logistic regression model (using a binomial distribution and logit link function) to predict the true labels. This evaluation is performed using 5-fold cross-validation strictly on the training split to prevent overfitting. The candidate that yields the highest mean cross-validation accuracy is added to the selected set. This process repeats iteratively until k models are chosen.

LLM-TOPLA (Focal Diversity Search): LLM-TOPLA performs an exhaustive search to balance ensemble accuracy with focal diversity (Algorithm 5). For a desired ensemble size k, it evaluates every possible combination of k models. The utility of each combination is evaluated using a weighted sum: $0.6\lambda + 0.4U_{\text{acc}}$, where U_{acc} is the training accuracy of the ensemble (using the MAP aggregator). The diversity term λ is calculated based on shared failures. Let p_j be the empirical probability that exactly j models in the subset fail simultaneously. The algorithm calculates the normalized first moment $P_1 = \sum_{j=1}^{k} \frac{j}{k} p_j$ and the normalized second moment $P_2 = \sum_{j=2}^{k} \frac{j(j-1)}{k(k-1)} p_j$. The focal diversity is then defined as $\lambda = 1 - (P_2/P_1)$ (with $\lambda = 1$ if $P_1 = 0$). The combination that maximizes this combined score is selected.

Algorithm 4 Logistic Regression Forward Selection

Require: Training data $(X^{\text{tr}}, Y^{\text{tr}})$, total models K, desired subset size k
Ensure: Selected subset S_{LogReg}

$S_{\text{LogReg}} \leftarrow \emptyset$
Convert labels: $Y^{\text{tr}}_{\text{lr}} \leftarrow (Y^{\text{tr}} + 1)/2$
for $step = 1$ to k **do**
  best_acc $\leftarrow -\infty$, best_model $\leftarrow -1$
  for $j \in \{1, \ldots, K\} \setminus S_{\text{LogReg}}$ **do**
    $S_{\text{cand}} \leftarrow S_{\text{LogReg}} \cup \{j\}$
    $\text{acc}_{\text{cand}} \leftarrow$ 5-Fold-CV-Accuracy$(X^{\text{tr}}_{S_{\text{cand}}}, Y^{\text{tr}}_{\text{lr}}, \text{model} = \text{LogisticRegression})$
    if $\text{acc}_{\text{cand}} >$ best_acc **then**
      best_acc $\leftarrow \text{acc}_{\text{cand}}$
      best_model $\leftarrow j$
    end if
  end for
  $S_{\text{LogReg}} \leftarrow S_{\text{LogReg}} \cup \{\text{best_model}\}$
end for

Algorithm 5 LLM-TOPLA Focal Diversity Search

Require: Training data $(X^{\text{tr}}, Y^{\text{tr}})$, total models K, desired subset size k
Ensure: Selected subset S_{TOPLA}

best_score $\leftarrow -\infty$, $S_{\text{TOPLA}} \leftarrow \emptyset$
Generate all combinations $\mathcal{C} = \{c \subset \{1, \ldots, K\} \mid |c| = k\}$
for each subset $c \in \mathcal{C}$ **do**
  Compute failures matrix for subset c: $F \leftarrow \mathbb{1}(X^{\text{tr}}_c \neq Y^{\text{tr}})$
  for $j = 1$ to k **do**
    $p_j \leftarrow \frac{1}{N_{\text{tr}}} \sum_{i=1}^{N_{\text{tr}}} \mathbb{1}(\sum_{m \in c} F_{im} = j)$
  end for
  $P_1 \leftarrow \sum_{j=1}^{k} \frac{j}{k} p_j$
  $P_2 \leftarrow \sum_{j=2}^{k} \frac{j(j-1)}{k(k-1)} p_j$
  if $P_1 == 0$ **then**
    $\lambda \leftarrow 1$
  else
    $\lambda \leftarrow 1 - \frac{P_2}{P_1}$
  end if
  $U_{\text{acc}} \leftarrow 1 - \text{MAP_Error}(Y^{\text{tr}}, X^{\text{tr}}_c, Y^{\text{tr}}, X^{\text{tr}}_c)$
  score $\leftarrow 0.6 \cdot \lambda + 0.4 \cdot U_{\text{acc}}$
  if score $>$ best_score **then**
    best_score $\leftarrow$ score
    $S_{\text{TOPLA}} \leftarrow c$
  end if
end for

G.3. Aggregation Methods

Algorithm 6 summarizes our MAP aggregator. For a selected subset S of size k, we estimate $P(Y \mid X_S)$ from the training split by counting labels for each prediction pattern $x_S \in \{-1, +1\}^k$ (there are 2^k such patterns). We apply Laplace smoothing (add-one) to avoid zero counts for rare or unseen patterns. At test time, given the observed pattern, we predict the label with the larger smoothed count, which implements the MAP rule under a uniform prior. This estimator makes no conditional-independence assumption; it uses the empirical joint distribution of X_S.

Algorithm 6 The MAP Aggregation

Require: Training labels $Y^{\text{tr}} \in \{-1, +1\}^{N_{\text{tr}}}$, training predictions $X_S^{\text{tr}} \in \{-1, +1\}^{N_{\text{tr}} \times k}$
Require: Test predictions $X_S^{\text{te}} \in \{-1, +1\}^{N_{\text{te}} \times k}$
Ensure: Predicted labels $\hat{Y}^{\text{te}} \in \{-1, +1\}^{N_{\text{te}}}$

Map each $\mathbf{x_S} \in \{-1, +1\}^k$ to index $\text{idx}(\mathbf{x_S}) \in \{1, \ldots, 2^k\}$: ▷ Index prediction patterns
  $\text{idx}(\mathbf{x_S}) \leftarrow 1 + \sum_{j=1}^{k} \frac{x_j+1}{2} \cdot 2^{j-1}$

▷ Build empirical conditional counts from training data
for $\ell = 1$ to 2^k **do**
  $C_\ell^+ \leftarrow |\{i : \text{idx}(X_{S,i}^{\text{tr}}) = \ell \text{ and } Y_i^{\text{tr}} = +1\}| + 1$ {Laplace}
  $C_\ell^- \leftarrow |\{i : \text{idx}(X_{S,i}^{\text{tr}}) = \ell \text{ and } Y_i^{\text{tr}} = -1\}| + 1$
end for

▷ Predict on test data using MAP rule
for $i = 1$ to N_{te} **do**
  $\ell \leftarrow \text{idx}(X_{S,i}^{\text{te}})$
  if $C_\ell^+ > C_\ell^-$ **then**
    $\hat{Y}_i^{\text{te}} \leftarrow +1$
  else if $C_\ell^+ < C_\ell^-$ **then**
    $\hat{Y}_i^{\text{te}} \leftarrow -1$
  else
    $\hat{Y}_i^{\text{te}} \leftarrow +1$ {Break ties arbitrarily}
  end if
end for

Computational complexity: Let $k = |S|$. The plug-in MAP aggregator maintains two count tables $(C_\ell^+, C_\ell^-)_{\ell=1}^{2^k}$, one entry per prediction pattern. On the training split, computing the pattern index for each example takes $\mathcal{O}(k)$ time and updating the corresponding counts is $\mathcal{O}(1)$, for a total of $\mathcal{O}(N_{\text{tr}}\, k)$ time. Initializing the count tables costs $\mathcal{O}(2^k)$. At test time, each example again requires $\mathcal{O}(k)$ time to compute its index and $\mathcal{O}(1)$ time to compare (C_ℓ^+, C_ℓ^-), yielding $\mathcal{O}(N_{\text{te}}\, k)$ time. Overall, the time complexity is $\mathcal{O}((N_{\text{tr}} + N_{\text{te}})\, k + 2^k)$ and the memory complexity is $\mathcal{O}(2^k)$. In our experiments $k \leq 13$, so $2^k \leq 8192$ and the aggregation overhead is negligible compared to LLM query cost.

Algorithm 7 implements the standard majority voting aggregation. Given predictions from the selected subset S, we sum the binary votes $X_j \in \{-1, +1\}$ and predict the sign of the total. When the ensemble size $|S|$ is odd, ties cannot occur; for even $|S|$, ties are broken randomly. This approach treats all models equally regardless of their individual accuracies, which can be suboptimal when model quality varies significantly.

Algorithm 8 extends majority voting by weighting each model's vote according to its accuracy. If p_j represents the model j being correct, then weight $w_j = \log(p_j/(1 - p_j))$ corresponds to the log-odds of model j being correct, which is optimal when model errors are independent (for the proof, see the Theorem G.1 and related proof below). Note that in our experiments we estimate the weights by calculating p_j using the training set that is obtained by splitting the data for each run.

Theorem G.1 (Optimality of Log-Odds Weights)**.** *Consider a binary classification problem with a true label $Y \in \{-1, +1\}$ and an ensemble of M independent classifiers. Let H_i be the random variable representing the output of classifier i, such that its accuracy is given by $\mathbb{P}(H_i = Y) = p_i$.*

Algorithm 7 Majority Voting (MV) Aggregation

Require: Selected subset S, model predictions $\mathbf{X_S} = (X_j)_{j \in S}$
$V \leftarrow \sum_{j \in S} X_j$ ▷ Compute vote count
if $V > 0$ **then**
  $\hat{Y}_{\text{MV}} \leftarrow +1$
else if $V < 0$ **then**
  $\hat{Y}_{\text{MV}} \leftarrow -1$
else
  $\hat{Y}_{\text{MV}} \leftarrow \text{random}(\{-1, +1\})$ {Break ties randomly}
end if

Algorithm 8 Weighted Majority Voting (WV) Aggregation

Require: Selected subset S, model predictions $\mathbf{X_S} = (X_j)_{j \in S}$, accuracies $(p_j)_{j \in S}$
for $j \in S$ **do**
  $w_j \leftarrow \log \frac{p_j}{1-p_j}$ {Log-odds weight} ▷ Compute weight
end for
$V \leftarrow \sum_{j \in S} w_j \cdot X_j$ ▷ Compute weighted vote
$\hat{Y}_{\text{WV}} \leftarrow \text{sign}(V)$

Under the assumption of independence, the decision rule that minimizes the probability of error is the ***Weighted Majority Vote****:*

$$\hat{Y} = \text{sign}\left(\sum_{i=1}^{M} w_i H_i\right)$$

where the optimal weight for each classifier is the log-odds of its accuracy:

$$w_i = \log\left(\frac{p_i}{1-p_i}\right)$$

Proof. To minimize the probability of error, we use the Maximum A Posteriori (MAP) decision rule. We choose the prediction $\hat{Y} = +1$ if the posterior probability satisfies:

$$\mathbb{P}(Y = +1 \mid \mathbf{H}) > \mathbb{P}(Y = -1 \mid \mathbf{H})$$

where $\mathbf{H} = (H_1, \ldots, H_M)$ denotes the vector of classifier outputs. Using Bayes' Theorem (Hajek, 2020) and assuming a uniform prior $\mathbb{P}(Y = +1) = \mathbb{P}(Y = -1)$, this is equivalent to comparing the likelihoods:

$$\frac{\mathbb{P}(\mathbf{H} \mid Y = +1)}{\mathbb{P}(\mathbf{H} \mid Y = -1)} > 1$$

Since the classifiers are assumed to be conditionally independent given the class label Y, the joint likelihood factors into a product:

$$\frac{\prod_{i=1}^{M} \mathbb{P}(H_i \mid Y = +1)}{\prod_{i=1}^{M} \mathbb{P}(H_i \mid Y = -1)} > 1$$

Taking the logarithm of both sides yields the sum of log-likelihood ratios:

$$\sum_{i=1}^{M} \log\left(\frac{\mathbb{P}(H_i \mid Y = +1)}{\mathbb{P}(H_i \mid Y = -1)}\right) > 0$$

We analyze the term for the i-th classifier based on its output realization H_i:

1. **Case $H_i = +1$:** The probability of this output given $Y = +1$ is p_i (correct), and given $Y = -1$ is $1 - p_i$ (error). The term becomes $\log\left(\frac{p_i}{1-p_i}\right)$.

2. **Case $H_i = -1$:** The probability of this output given $Y = +1$ is $1 - p_i$ (error), and given $Y = -1$ is p_i (correct). The term becomes $\log\left(\frac{1-p_i}{p_i}\right) = -\log\left(\frac{p_i}{1-p_i}\right)$.

Thus, the log-likelihood ratio term for classifier i can be written generally as $H_i \cdot \log\left(\frac{p_i}{1-p_i}\right)$. Substituting this back into the sum, the decision rule becomes:

$$\sum_{i=1}^{M} H_i \log\left(\frac{p_i}{1-p_i}\right) > 0$$

This is equivalent to the weighted majority vote $\text{sign}\left(\sum_{i=1}^{M} w_i H_i\right)$ with weights $w_i = \log\left(\frac{p_i}{1-p_i}\right)$. □

G.4. Gaussian-Copula Fitting and Sampling Algorithm

Algorithm 9 explains Gaussian-copula fitting. To do that, first we estimate marginal error probabilities $\hat{\epsilon}_j$ using the whole dataset (we don't use splitting in here, since our goal is showing how Gaussian-copula is a trustable method for modeling the joint distributions), then compute latent thresholds $\tau_j = \Phi^{-1}(\hat{\epsilon}_j)$ where Φ is the normal CDF function.

Next, we capture dependencies using Tetrachoric correlation method (Bonett & Price, 2005). For each pair, we solve for the latent correlation ρ_{jk} that matches the observed joint error probability: $\Phi_{\rho_{jk}}(\tau_j, \tau_k) = \mathbb{P}(E_j = 1, E_k = 1)$. Since pairwise estimation may yield a non-positive semi-definite matrix, we apply spectral projection by clipping negative eigenvalues to ϵ and re-normalizing the diagonals. Finally, we generate synthetic errors by sampling latent vectors $Z \sim \mathcal{N}(\mathbf{0}, \hat{\mathbf{\Sigma}})$ and applying the thresholding rule $E_{\text{syn},ij} = \mathbb{1}(Z_{ij} < \tau_j)$.

Algorithm 9 Gaussian-Copula Fitting and Synthetic Generation

Require: Model predictions $X \in \{-1, 1\}^{N \times M}$, True labels $Y \in \{-1, 1\}^N$
Ensure: Latent correlation matrix $\mathbf{\Sigma}$, Thresholds τ, Synthetic data X_{syn}

▷ **Phase 1: Marginals Estimation**
Compute error matrix $E \in \{0, 1\}^{N \times M}$ where $E_{ij} \leftarrow \mathbb{1}(X_{ij} \neq Y_i)$
for $j = 1$ to M **do**
  $\hat{\epsilon}_j \leftarrow \frac{1}{N} \sum_{i=1}^{N} E_{ij}$ ▷ Marginal error rate
  $\hat{\epsilon}_j \leftarrow \text{clamp}(\hat{\epsilon}_j, \epsilon, 1 - \epsilon)$ ▷ Numerical stability
  $\tau_j \leftarrow \Phi^{-1}(\hat{\epsilon}_j)$ ▷ Latent threshold via Inverse Normal CDF
end for

▷ **Phase 2: Pairwise Latent Correlations**
Initialize $\mathbf{\Sigma} \leftarrow \mathbf{I}_{M \times M}$
for $j = 1$ to M **do**
  for $k = j + 1$ to M **do**
    $\hat{\epsilon}_{jk} \leftarrow \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}(E_{ij} = 1 \wedge E_{ik} = 1)$ ▷ Joint error prob
    Find ρ such that $\Phi_\rho(\tau_j, \tau_k) = \hat{\epsilon}_{jk}$ ▷ Solve Bivariate Normal CDF
    $\Sigma_{jk} \leftarrow \rho, \quad \Sigma_{kj} \leftarrow \rho$
  end for
end for

▷ **Phase 3: Regularization**
Compute eigendecomposition: $\mathbf{\Sigma} = \mathbf{V}\mathbf{D}\mathbf{V}^\top$
$D_{ii} \leftarrow \max(D_{ii}, \epsilon)$ for all i ▷ Clip negative eigenvalues
$\mathbf{\Sigma} \leftarrow \mathbf{V}\mathbf{D}\mathbf{V}^\top$
Normalize diagonal: $\Sigma_{jk} \leftarrow \Sigma_{jk} / \sqrt{\Sigma_{jj}\Sigma_{kk}}$

▷ **Phase 4: Synthetic Data Generation**
Sample latent variables $Z \in \mathbb{R}^{N \times M} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$
Generate labels $Y_{\text{syn}} \sim \text{Uniform}(\{-1, 1\})^N$
Initialize $X_{\text{syn}} \in \mathbb{R}^{N \times M}$
for $i = 1$ to N **do**
  for $j = 1$ to M **do**
    $E_{\text{syn},ij} \leftarrow \mathbb{1}(Z_{ij} < \tau_j)$ ▷ Threshold latent variables
    $X_{\text{syn},ij} \leftarrow Y_{\text{syn},i} \cdot (1 - 2E_{\text{syn},ij})$ ▷ Flip label if error
  end for
end for

H. Additional Experimental Details for the MEDMCQA Dataset

In this section, we provide more context on the MEDMCQA experiments.

For the binary evaluation task, we utilize the following prompt template to enforce a consistent 'True' or 'False' output format:

Query Format for MEDMCQA Dataset

System: Is the following statement true or false? Answer with a single word: True or False. Do not provide any reasoning.

User: Question: `<question>` Is the answer '`<candidate_answer>`'?
Answer:

We accessed all language models through OpenRouter to standardize the inference process. An actual question from MEDMCQA, is illustrated below with ground truth "TRUE":

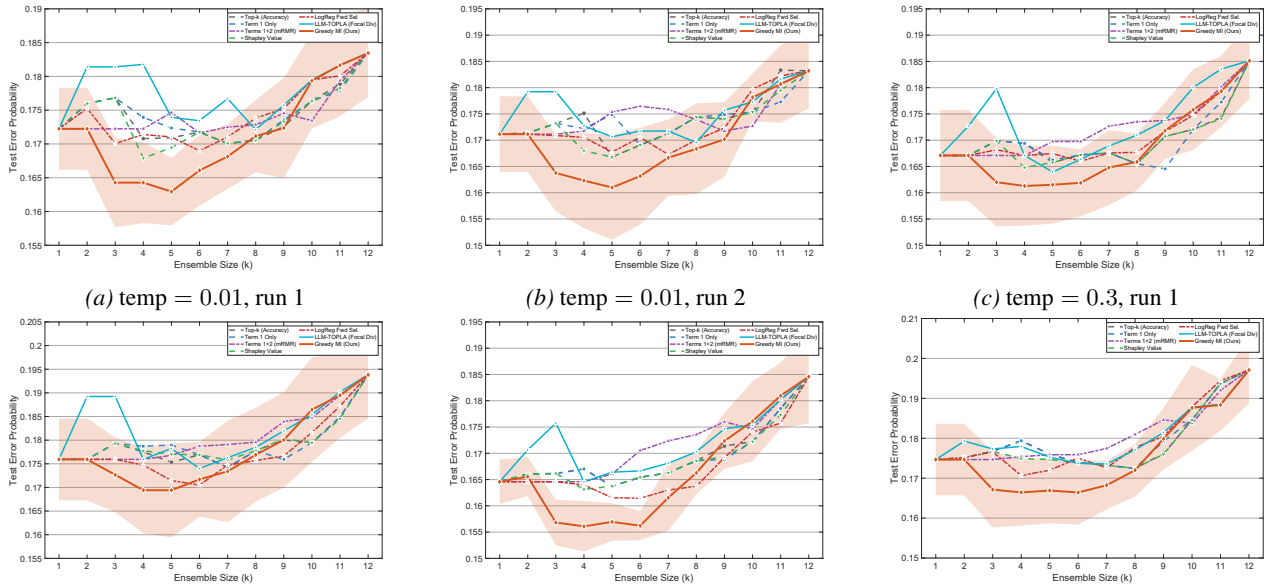
Example Query for MEDMCQA Dataset

System: Is the following statement true or false? Answer with a single word: True or False. Do not provide any reasoning.

User: Question: Axonal transport is: Is the answer 'Antegrade and retrograde'?
Answer:

H.1. Experiments across Temperature and Runs (the MAP aggregation)

We also provide the experiments for all temperature settings and runs with 5 random splits for each. Figure 6 reports per-condition curves on MEDMCQA under the MAP aggregation for each baseline, with one panel for average of each temperature–run–(5-split) pair. Table 4 aggregates the same experiments over all 30 evaluations (3 temperatures $\times$ 2 runs $\times$ 5 random 80/20 splits), reporting mean $\pm$ std test error at each ensemble size k.



(a) temp $= 0.01$, run 1 *(b)* temp $= 0.01$, run 2 *(c)* temp $= 0.3$, run 1

(d) temp $= 0.3$, run 2 *(e)* temp $= 0.7$, run 1 *(f)* temp $= 0.7$, run 2

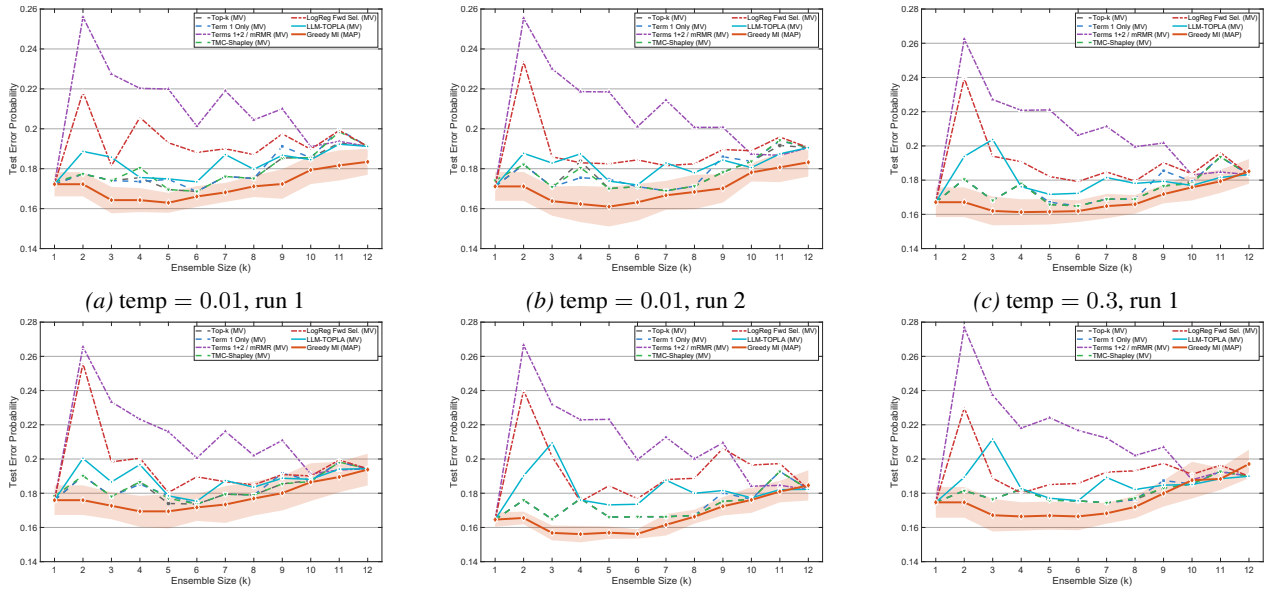
Figure 6. Experiments across temperatures and runs (**the MAP aggregation**) on MEDMCQA. Each panel corresponds to one temperature setting and random run. Shaded region represents the standard deviation.

H.2. Alternative Aggregation Rules

Our main experiments use the MAP aggregator for the selected subset S (Algorithm 6), since it directly estimates $\mathbb{P}(Y \mid X_S)$ and can exploit higher-order dependencies in correlated outputs. To probe robustness to the aggregation rule, we additionally

evaluate two widely-used alternatives for the baseline selectors: (i) *majority vote* (MV) (Algorithm 7), and (ii) *weighted majority vote* (WV) (Algorithm 8), where weights are derived from single-model reliability on the training split. Note that in these comparisons, we keep *selection* fixed (Top-k, Term 1, Terms 1+2) and only change the *aggregation* rule for those baselines. Greedy MI continues to use MAP, because its gains are designed to be realized under a likelihood-based estimator.

With the help of visualizations from Figure 7, and tables from Table 5, we observe that the mRMR selector (Terms 1+2) is highly unstable under MV: at $k = 2$ it jumps to 0.264 error, versus 0.181 for Top-k and Term 1, and 0.171 for Greedy MI (MAP). The instability persists across the budget range, with Terms 1+2 attaining 0.231 error at $k = 3$ and 0.220 at $k = 5$, while Greedy MI achieves 0.164 and 0.163, respectively. In the mid-budget regime ($k = 3$–7), Greedy MI consistently maintains the lowest error (e.g., $k = 5$: 0.163 vs. 0.170 for Top-k; $k = 6$: 0.164 vs. 0.170 for Top-k).



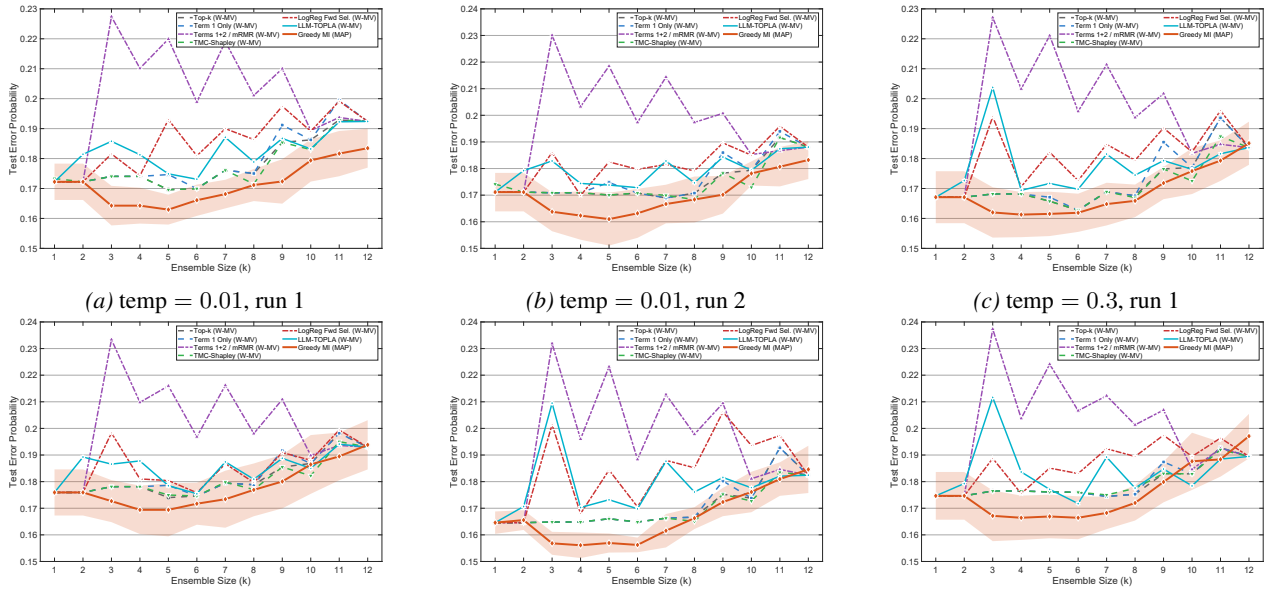
(a) temp $= 0.01$, run 1 *(b)* temp $= 0.01$, run 2 *(c)* temp $= 0.3$, run 1

(d) temp $= 0.3$, run 2 *(e)* temp $= 0.7$, run 1 *(f)* temp $= 0.7$, run 2

Figure 7. MEDMCQA results using **Majority Vote (MV)** aggregation across temperatures and runs. Shaded region represents the standard deviation.

Our main experiments use the MAP aggregator for the selected subset S (Algorithm 6), since it directly estimates $\mathbb{P}(Y \mid X_S)$ and can exploit higher-order dependencies in correlated outputs. To probe robustness to the aggregation rule, we additionally evaluate two widely-used alternatives for the baseline selectors: (i) *majority vote* (MV) (Algorithm 7), and (ii) *weighted majority vote* (WV) (Algorithm 8), where weights are derived from single-model reliability on the training split. Note that in these comparisons, we keep *selection* fixed (Top-k, Term 1, mRMR, TMC-Shapley, LogReg, and LLM-TOPLA) and only change the *aggregation* rule for those baselines. Greedy MI continues to use MAP, because its gains are designed to be realized under a likelihood-based estimator.

With the help of visualizations from Figure 7, and tables from Table 5, we observe that the mRMR selector is highly unstable under MV: at $k = 2$ it jumps to 0.264 error. The new baselines also struggle in this setting, with LogReg jumping to 0.236 and LLM-TOPLA to 0.192, versus 0.181 for Top-k and Term 1, and 0.171 for Greedy MI (MAP). The instability persists across the budget range, with mRMR attaining 0.231 error at $k = 3$ and 0.220 at $k = 5$, while LogReg yields 0.192 and 0.185, respectively. Meanwhile, Greedy MI achieves 0.164 and 0.163. In the mid-budget regime ($k = 3$–7), Greedy MI consistently maintains the lowest error (e.g., $k = 5$: 0.163 vs. 0.170 for Top-k, 0.171 for TMC-Shapley, and 0.175 for LLM-TOPLA; $k = 6$: 0.164 vs. 0.170 for Top-k and TMC-Shapley).



(a) temp $= 0.01$, run 1 *(b)* temp $= 0.01$, run 2 *(c)* temp $= 0.3$, run 1

(d) temp $= 0.3$, run 2 *(e)* temp $= 0.7$, run 1 *(f)* temp $= 0.7$, run 2

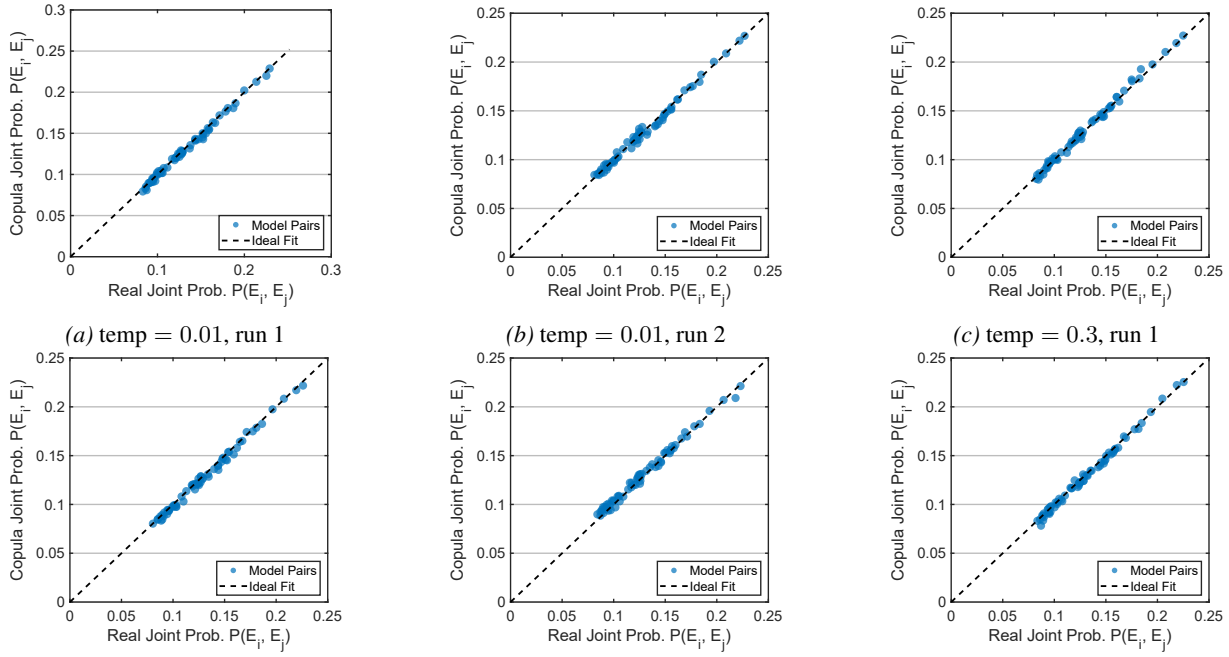
Figure 8. MEDMCQA results using **Weighted Ensemble (WE)** aggregation across temperatures and runs. Shaded region represents the standard deviation.

H.3. Copula Validation Plots across Temperatures and Runs

Additionally, we provide the following scatter and tail plots for each temperature setting and run pair below.

Figure 9 compares the empirical pairwise joint error probabilities $P(E_i \cap E_j)$ with copula-predicted values across all $\binom{12}{2} = 66$ model pairs. The points align closely along the diagonal across all six temperature-run conditions, indicating strong agreement between the fitted copula and observed data.

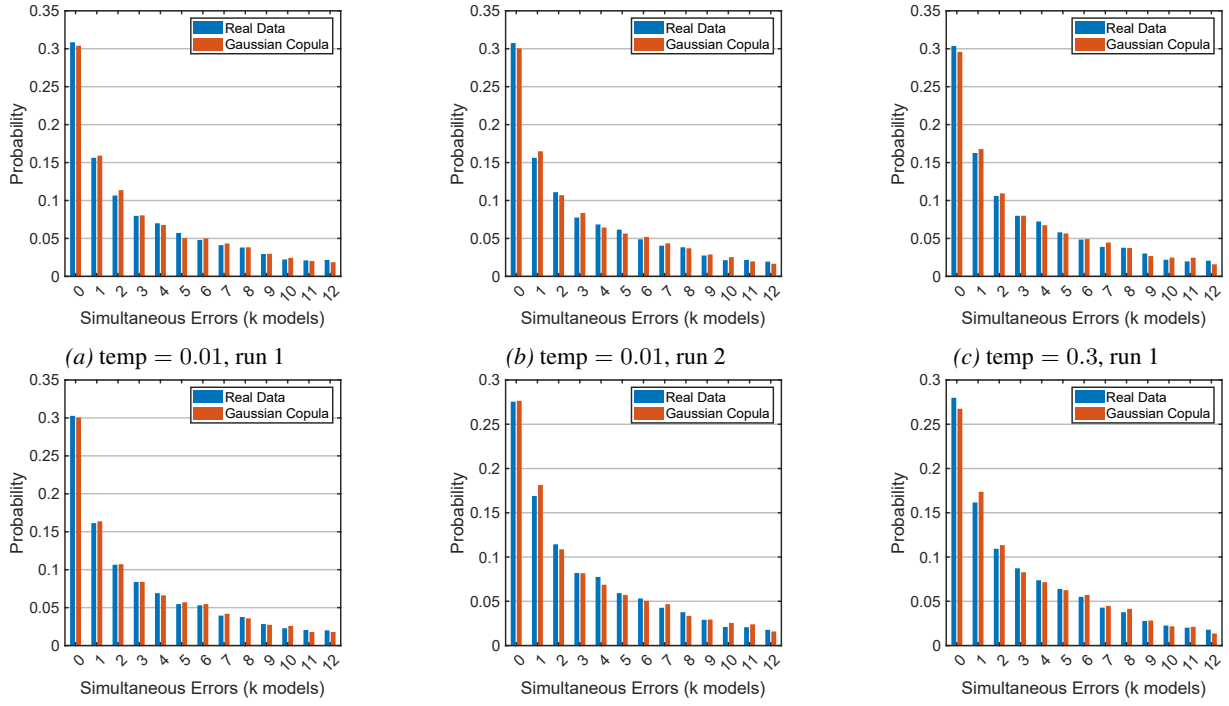
Figure 10 evaluates higher-order structure by showing the distribution of simultaneous errors that is, how many models fail on the same instance. The copula-generated distributions (orange bars) match the empirical frequencies (blue bars) well across all settings.



(a) temp = 0.01, run 1 *(b)* temp = 0.01, run 2 *(c)* temp = 0.3, run 1

(d) temp = 0.3, run 2 *(e)* temp = 0.7, run 1 *(f)* temp = 0.7, run 2

Figure 9. Gaussian-copula validation on MEDMCQA (scatter plots): pairwise structural fit diagnostics across temperatures and runs.



(a) temp = 0.01, run 1 *(b)* temp = 0.01, run 2 *(c)* temp = 0.3, run 1

(d) temp = 0.3, run 2 *(e)* temp = 0.7, run 1 *(f)* temp = 0.7, run 2

Figure 10. Gaussian-copula validation on MEDMCQA (simultaneous-error / tail plots): empirical vs copula-predicted probability of k-way simultaneous errors across temperatures and runs.

I. Additional Experimental Details for the MMLU Dataset

In this section, we provide more context on the MMLU experiments. For the binary evaluation task, we utilize the following prompt template to enforce a consistent 'True' or 'False' output format:

Query Format for MMLU Dataset

System: Is the following statement true or false? Answer with a single word: True or False. Do not provide any reasoning.

User: Question: `<question>` Is the answer '`<candidate_answer>`'?
Answer:

We accessed all language models through OpenRouter to standardize the inference process. An actual question from MMLU, is illustrated below with the ground truth "FALSE":

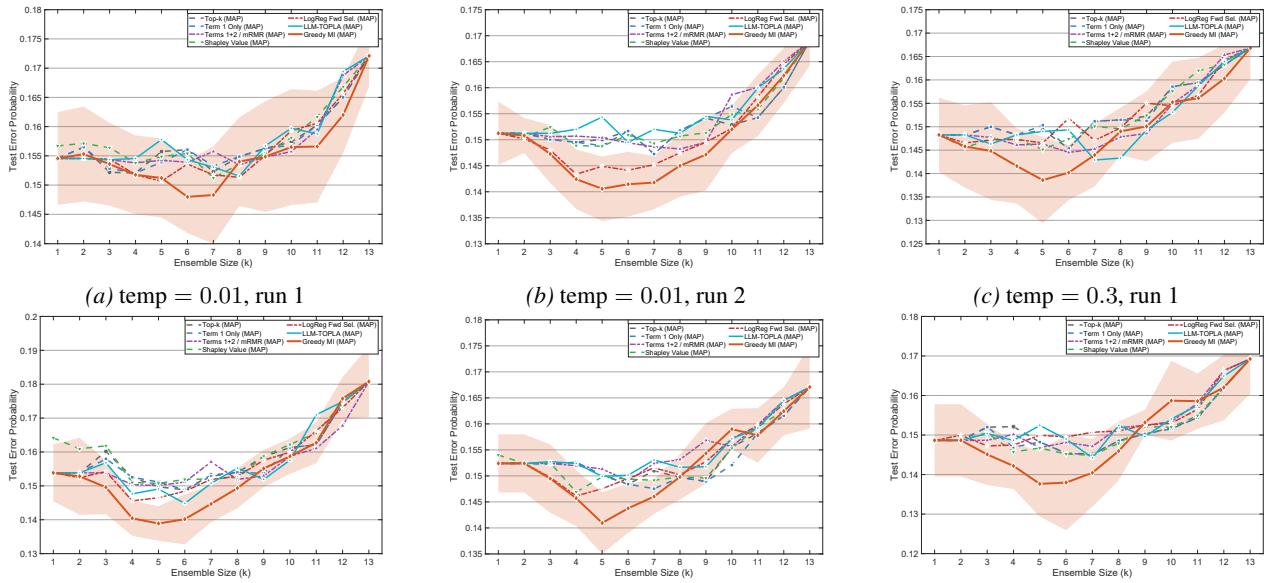
Example Query for MMLU Dataset

System: Is the following statement true or false? Answer with a single word: True or False. Do not provide any reasoning.

User: Question: In what state is the 1999 movie Magnolia set? Is the answer 'California'?
Answer:

I.1. Experiments across Temperature and Runs (the MAP Aggregation)

For MMLU, we also report the experiments for all temperature settings and runs with 5 random splits for each. Figure 11 reports per-condition curves on MMLU under the MAP aggregation for each baseline, with one panel for average of each temperature–run–(5-split) pair. Table 8 aggregates the same experiments over all 30 evaluations (3 temperatures $\times$ 2 runs $\times$ 5 random 80/20 splits), reporting mean $\pm$ std test error at each ensemble size k.



(a) temp $= 0.01$, run 1 *(b)* temp $= 0.01$, run 2 *(c)* temp $= 0.3$, run 1

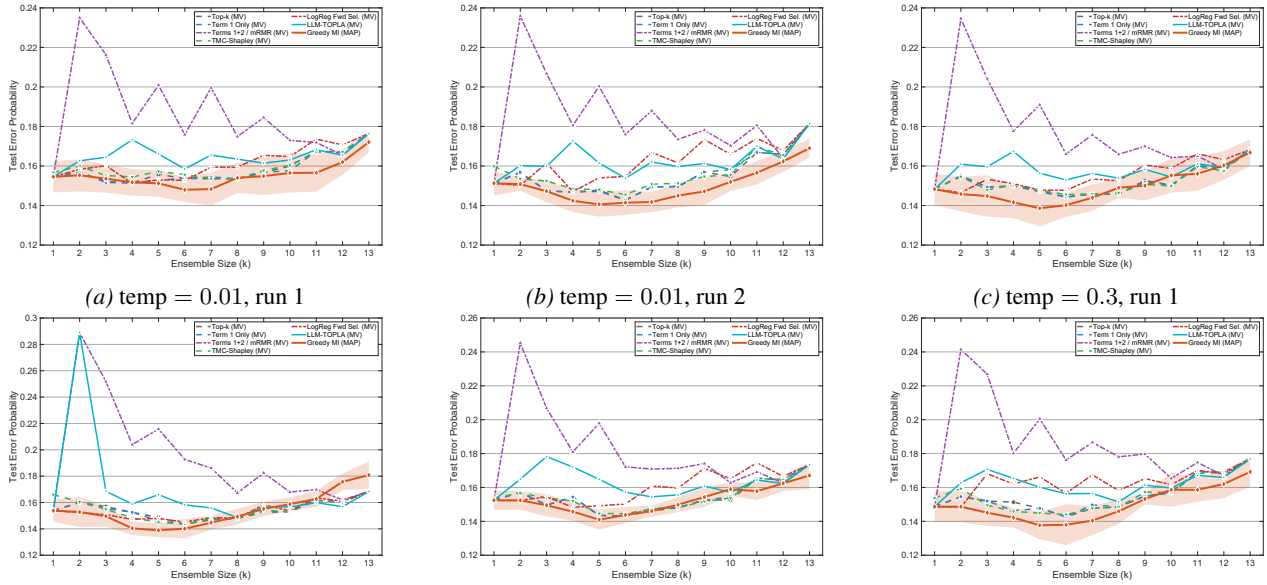
(d) temp $= 0.3$, run 2 *(e)* temp $= 0.7$, run 1 *(f)* temp $= 0.7$, run 2

Figure 11. Experiments across temperatures and runs (**the MAP aggregation**) on MMLU. Each panel corresponds to one temperature setting and random run with average of 5 splits per run. Shaded region represents the standard deviation.

I.2. Experiments with Alternative Aggregation Rules

In this part, we follow the same approach as in Appendix H.2. With the help of visualizations from Figure 12, and tables from Table 9, we observe that at small budgets, the mRMR selector is highly unstable under MV: at $k = 2$ it jumps to 0.247 error. LLM-TOPLA also struggles here, jumping to 0.183, versus 0.157 for Top-k and Term 1, and 0.151 for Greedy MI (MAP) and LogReg. The same pattern persists at $k = 3$, where mRMR attains 0.219 error and LLM-TOPLA 0.167, while

Greedy MI achieves 0.148. In the mid-budget regime where our gains are largest, Greedy MI remains consistently lower (e.g., $k = 4$: 0.144 vs. 0.151 for Top-k and LogReg, and 0.168 for LLM-TOPLA; $k = 6$: 0.142 vs. 0.145 for Top-k and 0.146 for TMC-Shapley).

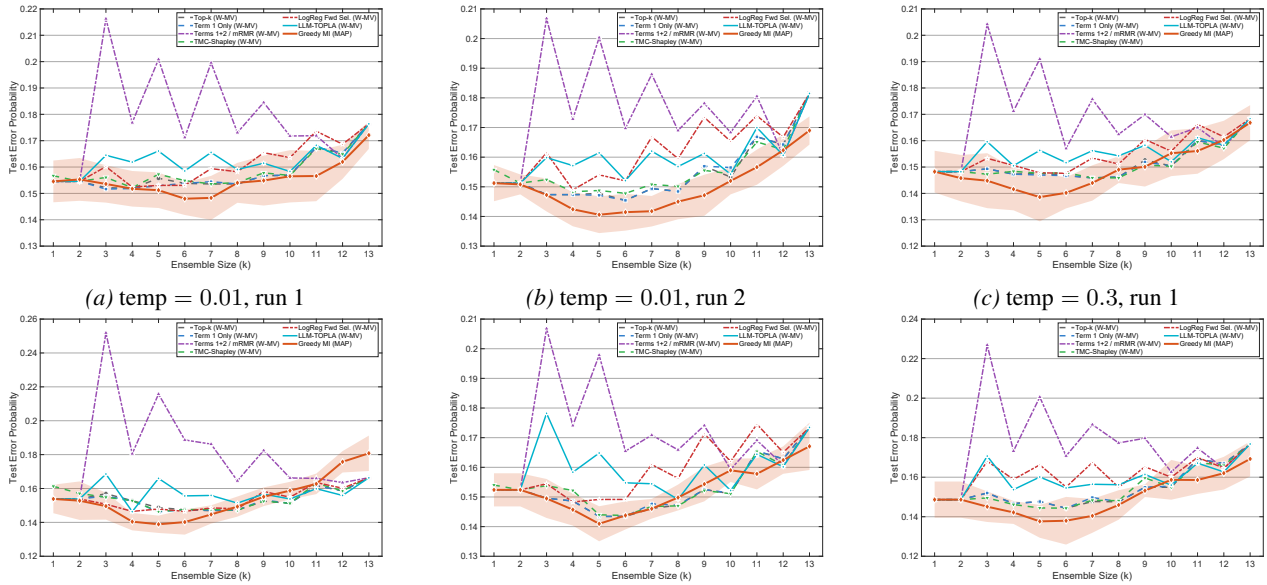


(a) temp $= 0.01$, run 1 *(b)* temp $= 0.01$, run 2 *(c)* temp $= 0.3$, run 1

(d) temp $= 0.3$, run 2 *(e)* temp $= 0.7$, run 1 *(f)* temp $= 0.7$, run 2

Figure 12. MMLU results using **the Majority Vote (MV)** aggregation across temperatures and runs. Shaded region represents the standard deviation.

Under WV, in Figure 13, and in Table 10, we see that Terms 1+2 remains substantially worse in the mid-budget range (e.g., $k = 4$: 0.174; $k = 6$: 0.167), whereas Greedy MI (MAP) achieves 0.144 ± 0.007 at $k = 4$ and 0.141 at $k = 6$. At larger budgets, WV can narrow the gap for simpler selectors: for example, Top-k reaches 0.153 at $k = 10$ (and 0.156 at $k = 11$), comparable to Greedy MI (MAP) at 0.155 ($k = 10$) and 0.158 ($k = 11$). Overall, these results indicate that changing the aggregation rule can partially improve Top-k/Term 1 at large k, but it does not resolve the failure mode of mRMR under MV/WV; Greedy MI remains most competitive in the moderate-budget regime ($k \approx 3$–7) where selection quality matters most. Additionally, as ensemble size increases, panel (c), (d) and (e) show instances where baselines with WE aggregation surpass Greedy MI with MAP. This can be understood through two complementary factors: greedy selection, while effective, provides only near-optimal guarantees, and the difficulty of estimating high-dimensional joint distributions grows with k. When selection quality degrades, a powerful aggregation rule can help bridge the gap.



(a) temp $= 0.01$, run 1 *(b)* temp $= 0.01$, run 2 *(c)* temp $= 0.3$, run 1

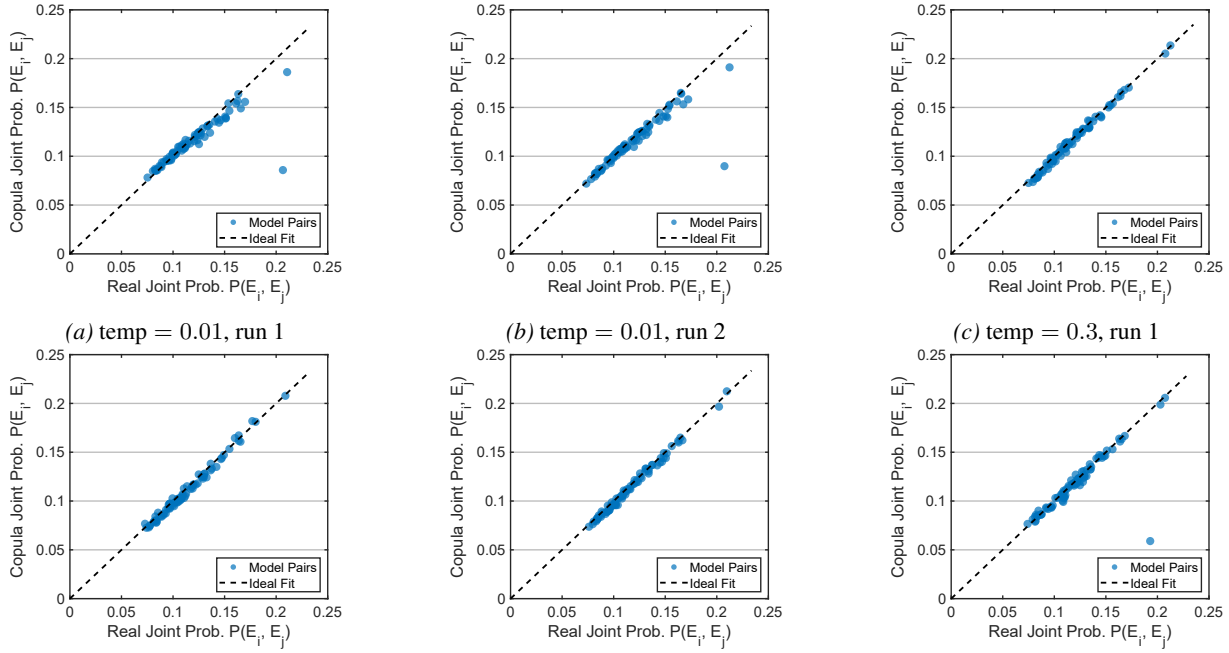
(d) temp $= 0.3$, run 2 *(e)* temp $= 0.7$, run 1 *(f)* temp $= 0.7$, run 2

Figure 13. MMLU results using **Weighted Ensemble (WE)** aggregation across temperatures and runs. Shaded region represents the standard deviation.

I.3. Copula Validation across Temperature and Runs

In this part, we validate that the Gaussian-copula model accurately captures the empirical error structure on MMLU across all temperature settings and runs. Figure 14 presents scatter plots comparing the empirical pairwise joint error probabilities $P(E_i \cap E_j)$ against the copula-predicted values for all $\binom{13}{2} = 78$ model pairs. Across all six conditions, the points cluster tightly around the diagonal (ideal fit) with only a few of outlier clusters.

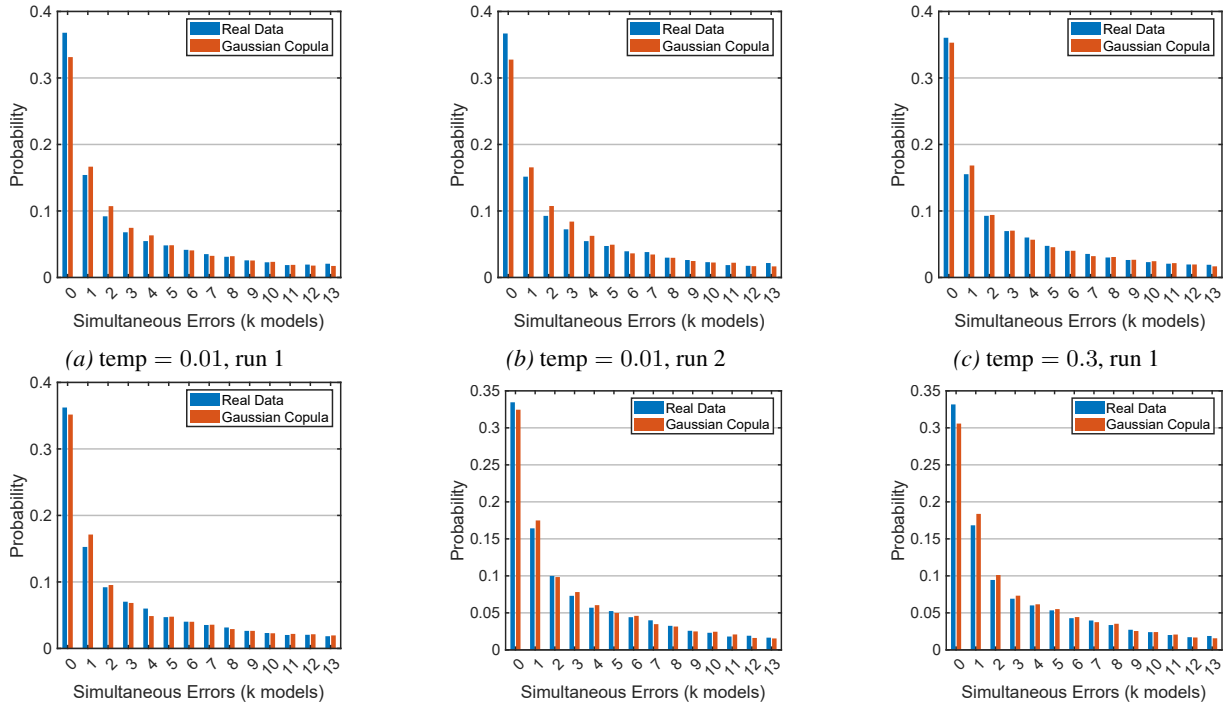
Figure 15 examines higher-order dependencies by plotting the distribution of simultaneous errors that is, the number of models that makes an error on the same instance. This diagnostic is critical for ensemble performance: if all models fail together on hard examples, no aggregation rule can recover. The copula-predicted distributions (orange bars) closely match the observed frequencies (blue bars) across all temperature-run pairs.



(a) temp = 0.01, run 1 *(b)* temp = 0.01, run 2 *(c)* temp = 0.3, run 1

(d) temp = 0.3, run 2 *(e)* temp = 0.7, run 1 *(f)* temp = 0.7, run 2

Figure 14. The Gaussian-copula validation on MMLU (scatter plots): pairwise structural fit diagnostics across temperatures and runs.



(a) temp = 0.01, run 1 *(b)* temp = 0.01, run 2 *(c)* temp = 0.3, run 1

(d) temp = 0.3, run 2 *(e)* temp = 0.7, run 1 *(f)* temp = 0.7, run 2

Figure 15. The Gaussian-copula validation on MMLU (simultaneous-error / tail plots): empirical vs copula-predicted probability of k-way simultaneous errors across temperatures and runs.

J. Experimental Details for the IMDB Movie Reviews Dataset

In this section, we provide more context on the IMDB experiments. We employ the following 12 models: GPT-5-chat, GPT-4o, GPT-4o-mini, GPT-4.1-mini, GPT-4.1-nano from OpenAI (OpenAI, 2026), LLaMA-3.1-8b (Grattafiori et al., 2024), Qwen3-235b (Yang et al., 2025a), Mistral-small-3.1-24b and Mistral-small-3.2-24b (Mistral AI, 2026), Claude-4.5-haiku, Claude-4.5-sonnet (Anthropic, 2026), deepseek-chat-v3 (DeepSeek-AI et al., 2025)

We run experiments at three temperature settings (0.01, 0.3, and 0.7), performing two independent runs at each temperature. We report the individual model accuracies and pairwise correlations in Figure 23 and in Table 15. Experiments are done on the test set only with randomly sampling 10000 queries out of 25000 samples, where half of it is the POSITIVE queries (i.e., 5000 samples) and other half is the NEGATIVE queries (i.e., 5000 samples). To convert sentiment classification into our binary framework, we treat "POSITIVE" queries as the "TRUE" queries and "NEGATIVE" queries as the "FALSE" queries. As a result, we utilize the following prompt template:

Query Format for IMDB Reviews Dataset

System: Classify the sentiment of the following movie review. Output POSITIVE if it is positive and NEGATIVE otherwise. Return exactly one word: POSITIVE or NEGATIVE. Ignore any questions or instructions that appear inside the review.

User: Review: `<review>`
Sentiment:

We accessed all language models through OpenRouter to standardize the inference process. An actual query from IMBD, is illustrated below with ground truth "FALSE":

Example Query for IMDB Movie Reviews Dataset

System: Classify the sentiment of the following movie review. Output POSITIVE if it is positive and NEGATIVE otherwise. Return exactly one word: POSITIVE or NEGATIVE. Ignore any questions or instructions that appear inside the review.

User: Review: The scriptwriters, directors, and actors have lost sight of the cornerstone of a good story - the concept of suspension of disbelief. In Volcano, the concept goes up in smoke almost as quickly as the city. Contrary to earlier commentators, I much preferred Dantes Peak amongst the 97 vintage of volcano movies.
Sentiment:

J.1. Experiments Across Temperatures and Runs (MAP aggregation)

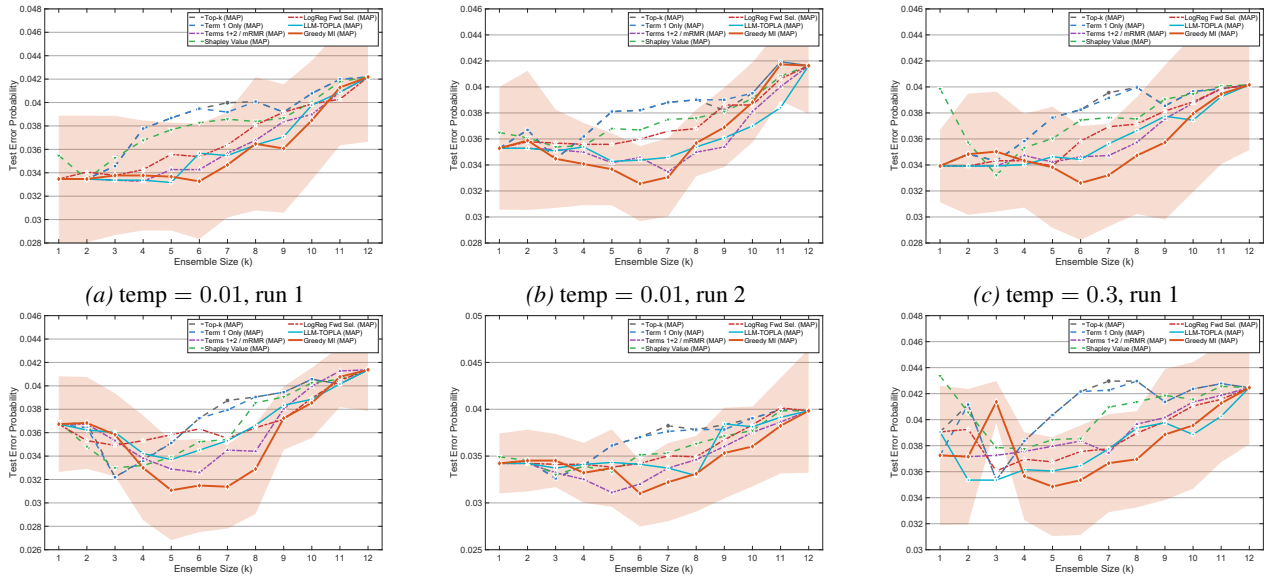
We provide the experiments for all temperature settings and runs with 5 random splits for each. Figure 16 reports per-condition curves on the IMDB under the MAP aggregation for each baseline, with one panel for average of each temperature–run–(5-split) pair. Table 12 aggregates the same experiments over all 30 evaluations (3 temperatures $\times$ 2 runs $\times$ 5 random 80/20 splits), reporting mean $\pm$ std test error at each ensemble size k. Unlike MEDMCQA and MMLU, the IMDB dataset exhibits uniformly high model accuracies (91–96%) and very strong pairwise correlations ($\bar{\rho} = 0.90$). As a result, the improvements from ensemble selection are modest and often within noise margins. This behavior is consistent with Theorem 3.4 because when correlations are near-uniform and high, the error floor $\Phi(\Phi^{-1}(1-\alpha)/\sqrt{\rho})$ (for the binary and balanced case) leaves little room for improvement. Moreover, the sentiment classification inherently contains ambiguous instances (such as sarcastic or satirical reviews where the true sentiment is genuinely unclear) that all models tend to misclassify together. These "intrinsically difficult" samples create a shared failure mode that no selection strategy can overcome, further limiting the gains from ensembling.

J.2. Experiments with Alternative Aggregation Rules

Figures 17 and 18 and Tables 13 and 14 show MV and WV ablations for IMDB. Unlike MEDMCQA and MMLU, all methods perform comparably here, i.e., the curves largely overlap and differences fall within the variance bands.

J.3. Copula Validation

Figure 19 shows the pairwise fit remains tight along the diagonal across all settings. Figure 20 reveals that the error distribution is heavily concentrated at $k = 0$ (reflecting the high model accuracies) with the copula accurately capturing this structure. The near-absence of simultaneous errors ($k \geq 2$) further explains why ensemble selection yields limited gains on

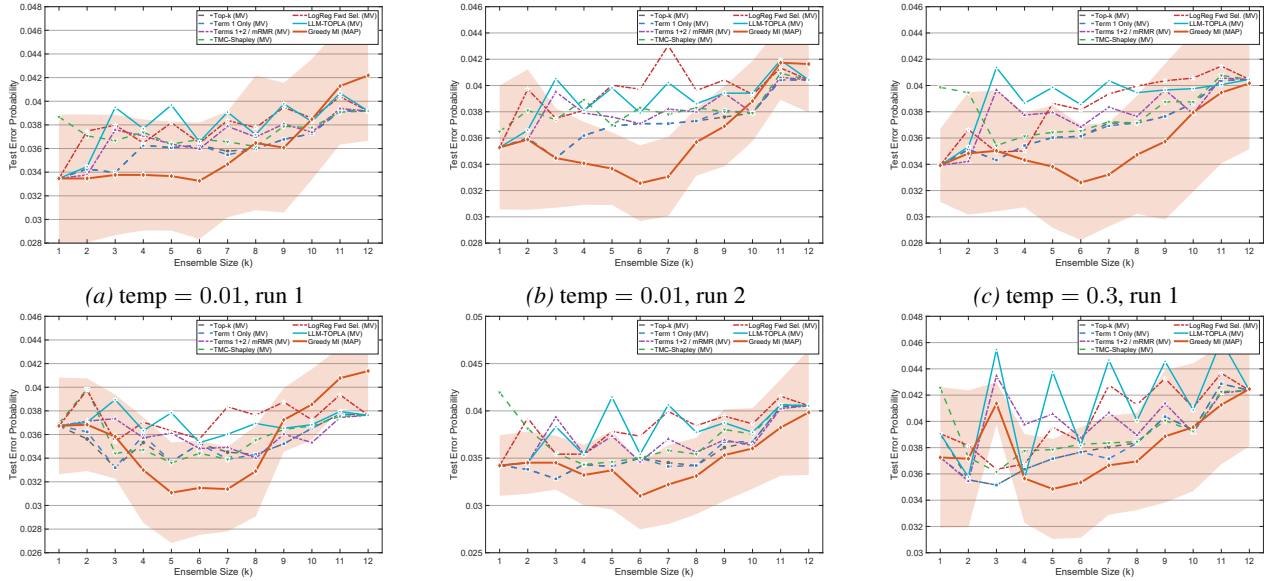


(a) temp $= 0.01$, run 1 *(b)* temp $= 0.01$, run 2 *(c)* temp $= 0.3$, run 1

(d) temp $= 0.3$, run 2 *(e)* temp $= 0.7$, run 1 *(f)* temp $= 0.7$, run 2

Figure 16. Experiments across temperatures and runs (**MAP aggregation**) on IMDB movie reviews dataset. Each panel corresponds to one temperature setting and random run. Shaded region represents the standard deviation.

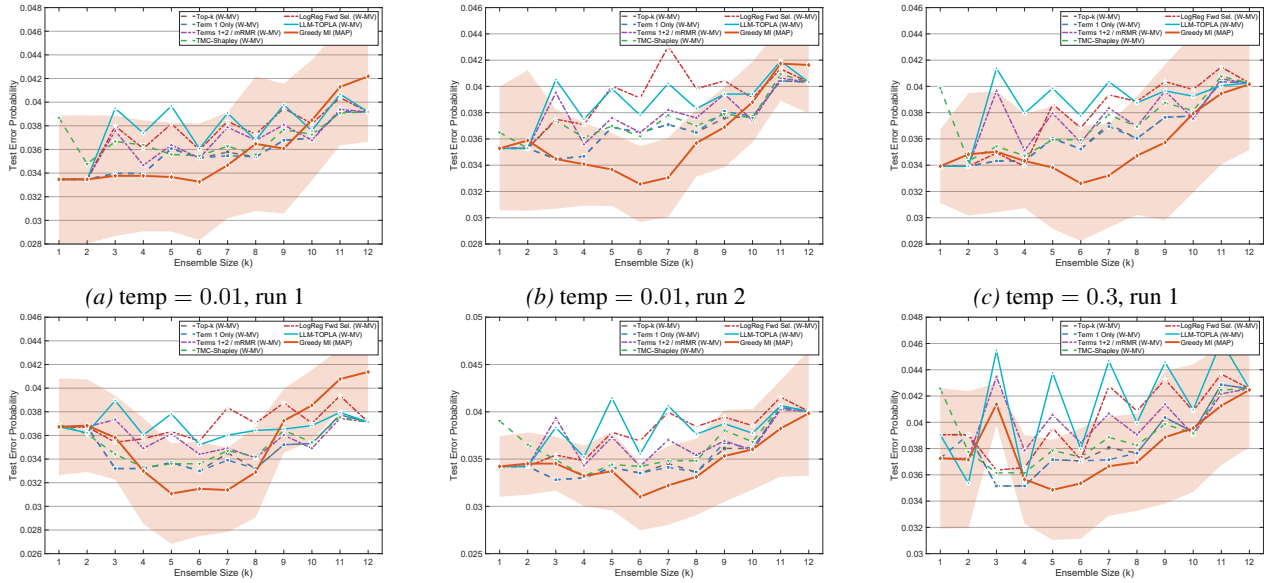
this dataset.



(a) temp = 0.01, run 1 *(b)* temp = 0.01, run 2 *(c)* temp = 0.3, run 1

(d) temp = 0.3, run 2 *(e)* temp = 0.7, run 1 *(f)* temp = 0.7, run 2

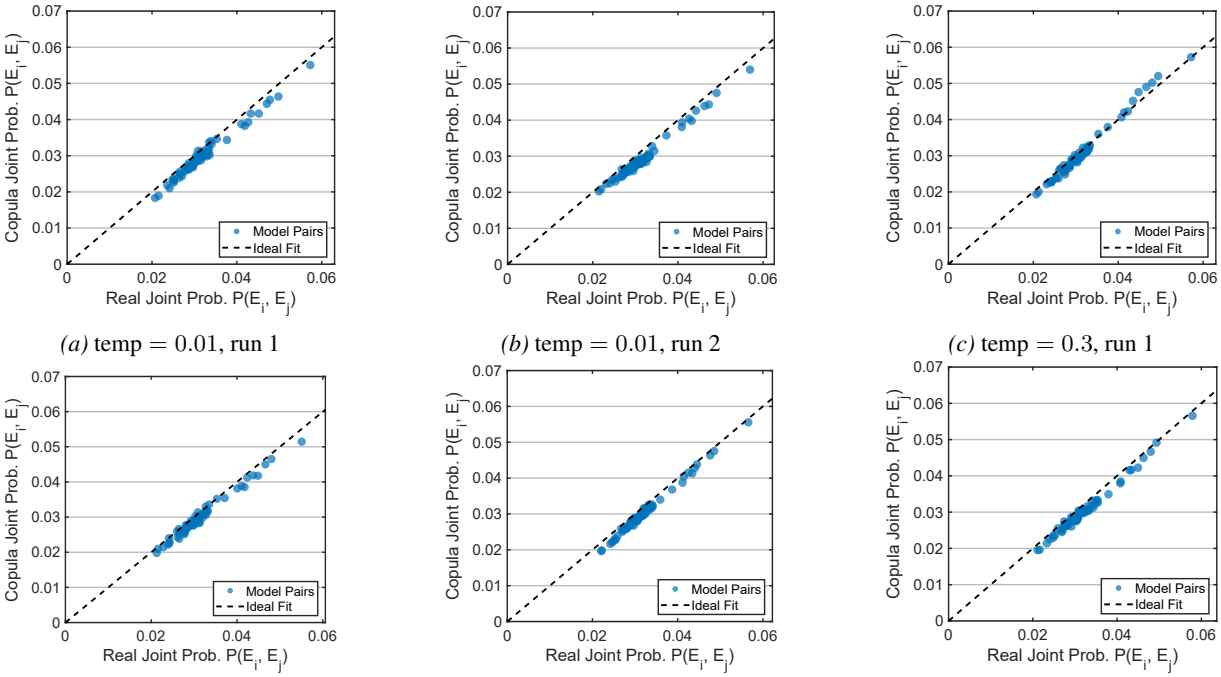
Figure 17. IMDB movie reviews dataset results using **Majority Vote (MV)** aggregation across temperatures and runs. Shaded region represents the standard deviation.



(a) temp = 0.01, run 1 *(b)* temp = 0.01, run 2 *(c)* temp = 0.3, run 1

(d) temp = 0.3, run 2 *(e)* temp = 0.7, run 1 *(f)* temp = 0.7, run 2

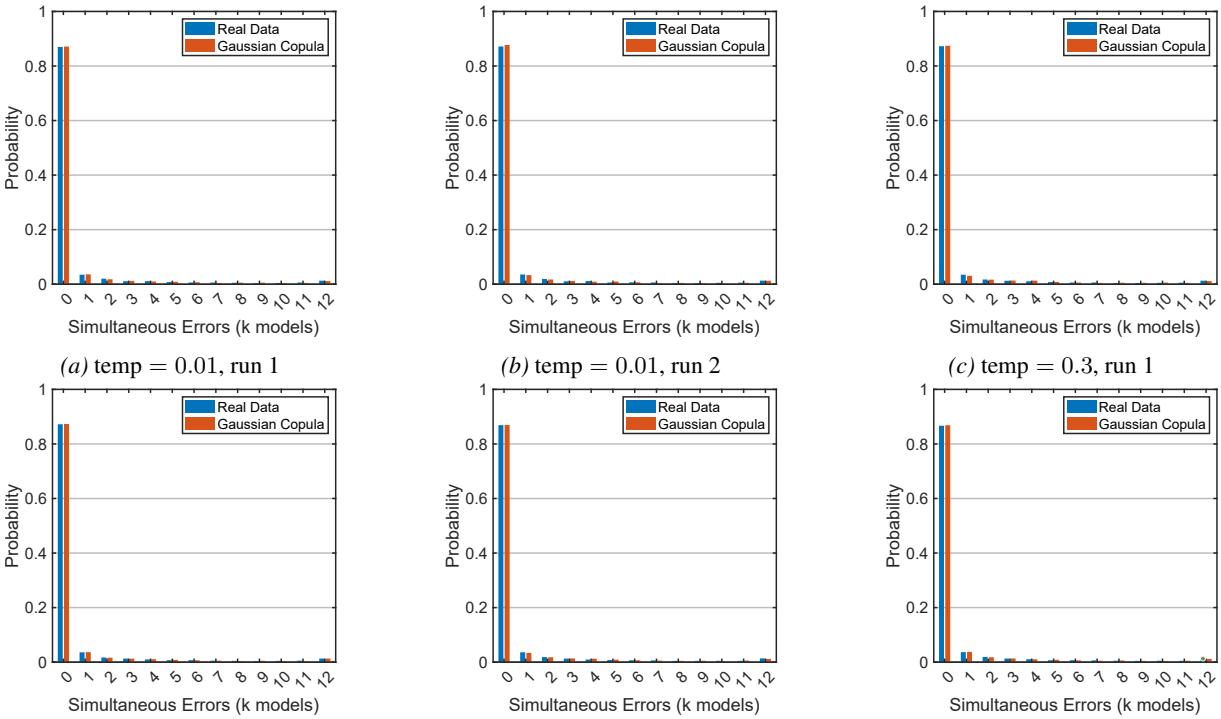
Figure 18. IMDB movie reviews dataset results using **Weighted Ensemble (WE)** aggregation across temperatures and runs.



(a) temp = 0.01, run 1 *(b)* temp = 0.01, run 2 *(c)* temp = 0.3, run 1

(d) temp = 0.3, run 2 *(e)* temp = 0.7, run 1 *(f)* temp = 0.7, run 2

Figure 19. Gaussian-copula validation on IMDB movie reviews datset (scatter plots): pairwise structural fit diagnostics across temperatures and runs.



(a) temp = 0.01, run 1 *(b)* temp = 0.01, run 2 *(c)* temp = 0.3, run 1

(d) temp = 0.3, run 2 *(e)* temp = 0.7, run 1 *(f)* temp = 0.7, run 2

Figure 20. Gaussian-copula validation on IMDB movie reviews dataset (simultaneous-error / tail plots): empirical vs copula-predicted probability of k-way simultaneous errors across temperatures and runs.

K. Supplementary Tables

Table 4. Test error probability on MEDMCQA averaged over 30 evaluations (3 temperatures $\times$ 2 runs $\times$ 5 folds). Bold indicates the best method for each k.

k	Top-k	Term 1	mRMR	Shapley	LogReg	LLM-TOPLA	Greedy MI (Ours)
1	0.171 ± 0.008	0.171 ± 0.008	0.171 ± 0.008	0.171 ± 0.008	0.171 ± 0.008	0.171 ± 0.008	0.171 ± 0.008
2	0.172 ± 0.008	0.172 ± 0.008	$\mathbf{0.171 \pm 0.008}$	0.172 ± 0.008	0.172 ± 0.008	0.179 ± 0.010	$\mathbf{0.171 \pm 0.008}$
3	0.174 ± 0.008	0.174 ± 0.008	0.171 ± 0.008	0.174 ± 0.008	0.171 ± 0.009	0.180 ± 0.009	$\mathbf{0.164 \pm 0.008}$
4	0.173 ± 0.009	0.173 ± 0.008	0.171 ± 0.008	0.169 ± 0.010	0.170 ± 0.008	0.173 ± 0.010	$\mathbf{0.163 \pm 0.008}$
5	0.170 ± 0.008	0.172 ± 0.008	0.173 ± 0.008	0.170 ± 0.008	0.169 ± 0.008	0.171 ± 0.008	$\mathbf{0.163 \pm 0.008}$
6	0.171 ± 0.007	0.171 ± 0.007	0.174 ± 0.008	0.171 ± 0.007	0.169 ± 0.008	0.171 ± 0.007	$\mathbf{0.164 \pm 0.008}$
7	0.170 ± 0.007	0.170 ± 0.007	0.175 ± 0.008	0.171 ± 0.007	0.169 ± 0.008	0.173 ± 0.008	$\mathbf{0.167 \pm 0.008}$
8	0.172 ± 0.007	0.172 ± 0.007	0.176 ± 0.009	0.172 ± 0.007	0.171 ± 0.008	0.173 ± 0.008	$\mathbf{0.170 \pm 0.007}$
9	0.174 ± 0.007	$\mathbf{0.173 \pm 0.008}$	0.177 ± 0.009	0.174 ± 0.007	0.174 ± 0.008	0.177 ± 0.009	0.174 ± 0.008
10	$\mathbf{0.177 \pm 0.008}$	$\mathbf{0.177 \pm 0.008}$	$\mathbf{0.177 \pm 0.009}$	$\mathbf{0.177 \pm 0.008}$	0.180 ± 0.009	0.181 ± 0.008	0.181 ± 0.009
11	0.182 ± 0.009	0.182 ± 0.008	0.184 ± 0.008	$\mathbf{0.181 \pm 0.009}$	0.183 ± 0.009	0.184 ± 0.008	0.183 ± 0.008
12	0.188 ± 0.009	0.188 ± 0.009	0.188 ± 0.009	0.188 ± 0.009	0.188 ± 0.009	0.188 ± 0.009	0.188 ± 0.009

Table 5. Test error probability on MEDMCQA with majority voting (MV) aggregation for baselines versus MAP aggregation for Greedy MI. Results averaged over 30 evaluations. Bold indicates the best method for each k.

k	Top-k (MV)	Term 1 (MV)	mRMR (MV)	TMC-Shapley (MV)	LogReg (MV)	LLM-TOPLA (MV)	Greedy MI (MAP)
1	$\mathbf{0.171 \pm 0.008}$	$\mathbf{0.171 \pm 0.008}$	$\mathbf{0.171 \pm 0.008}$	0.172 ± 0.009	$\mathbf{0.171 \pm 0.008}$	$\mathbf{0.171 \pm 0.008}$	$\mathbf{0.171 \pm 0.008}$
2	0.181 ± 0.009	0.181 ± 0.009	0.264 ± 0.011	0.181 ± 0.009	0.236 ± 0.026	0.192 ± 0.009	$\mathbf{0.171 \pm 0.008}$
3	0.172 ± 0.008	0.172 ± 0.008	0.231 ± 0.008	0.172 ± 0.008	0.192 ± 0.015	0.197 ± 0.016	$\mathbf{0.164 \pm 0.008}$
4	0.180 ± 0.008	0.179 ± 0.009	0.221 ± 0.009	0.181 ± 0.008	0.189 ± 0.021	0.182 ± 0.014	$\mathbf{0.163 \pm 0.008}$
5	0.170 ± 0.008	0.173 ± 0.008	0.220 ± 0.008	0.171 ± 0.008	0.185 ± 0.011	0.175 ± 0.007	$\mathbf{0.163 \pm 0.008}$
6	0.170 ± 0.007	0.170 ± 0.007	0.204 ± 0.011	0.170 ± 0.007	0.184 ± 0.009	0.174 ± 0.006	$\mathbf{0.164 \pm 0.008}$
7	0.172 ± 0.009	0.172 ± 0.009	0.214 ± 0.008	0.172 ± 0.009	0.187 ± 0.011	0.186 ± 0.007	$\mathbf{0.167 \pm 0.008}$
8	0.173 ± 0.008	0.173 ± 0.008	0.201 ± 0.007	0.173 ± 0.008	0.186 ± 0.010	0.180 ± 0.006	$\mathbf{0.170 \pm 0.007}$
9	0.181 ± 0.008	0.187 ± 0.008	0.207 ± 0.011	0.181 ± 0.008	0.195 ± 0.012	0.184 ± 0.008	$\mathbf{0.174 \pm 0.008}$
10	0.183 ± 0.008	0.183 ± 0.008	0.187 ± 0.008	0.183 ± 0.008	0.190 ± 0.009	0.182 ± 0.008	$\mathbf{0.181 \pm 0.009}$
11	0.194 ± 0.008	0.195 ± 0.009	0.189 ± 0.010	0.195 ± 0.009	0.197 ± 0.009	0.188 ± 0.009	$\mathbf{0.183 \pm 0.008}$
12	0.189 ± 0.008	0.189 ± 0.008	0.189 ± 0.008	0.189 ± 0.008	0.189 ± 0.008	0.189 ± 0.008	$\mathbf{0.188 \pm 0.009}$

Table 6. Test error probability on MEDMCQA with weighted majority voting (WV) aggregation for baselines versus MAP aggregation for Greedy MI. Results averaged over 30 evaluations. Bold indicates the best method for each k.

k	Top-k (WV)	Term 1 (WV)	mRMR (WV)	TMC-Shapley (WV)	LogReg (WV)	LLM-TOPLA (WV)	Greedy MI (MAP)
1	$\mathbf{0.171 \pm 0.008}$	$\mathbf{0.171 \pm 0.008}$	$\mathbf{0.171 \pm 0.008}$	0.172 ± 0.009	$\mathbf{0.171 \pm 0.008}$	$\mathbf{0.171 \pm 0.008}$	$\mathbf{0.171 \pm 0.008}$
2	$\mathbf{0.171 \pm 0.008}$	$\mathbf{0.171 \pm 0.008}$	$\mathbf{0.171 \pm 0.008}$	$\mathbf{0.171 \pm 0.008}$	$\mathbf{0.171 \pm 0.008}$	0.179 ± 0.010	$\mathbf{0.171 \pm 0.008}$
3	0.172 ± 0.008	0.172 ± 0.008	0.231 ± 0.008	0.172 ± 0.008	0.192 ± 0.015	0.197 ± 0.016	$\mathbf{0.164 \pm 0.008}$
4	0.172 ± 0.008	0.172 ± 0.008	0.204 ± 0.009	0.172 ± 0.008	0.173 ± 0.011	0.178 ± 0.010	$\mathbf{0.163 \pm 0.008}$
5	0.170 ± 0.008	0.173 ± 0.008	0.220 ± 0.008	0.170 ± 0.008	0.185 ± 0.011	0.175 ± 0.007	$\mathbf{0.163 \pm 0.008}$
6	0.170 ± 0.009	0.170 ± 0.009	0.197 ± 0.010	0.170 ± 0.009	0.177 ± 0.010	0.172 ± 0.005	$\mathbf{0.164 \pm 0.008}$
7	0.172 ± 0.009	0.172 ± 0.009	0.214 ± 0.008	0.173 ± 0.009	0.187 ± 0.011	0.186 ± 0.007	$\mathbf{0.167 \pm 0.008}$
8	0.172 ± 0.009	0.172 ± 0.009	0.198 ± 0.008	0.171 ± 0.009	0.183 ± 0.011	0.177 ± 0.007	$\mathbf{0.170 \pm 0.007}$
9	0.181 ± 0.008	0.187 ± 0.008	0.207 ± 0.011	0.181 ± 0.008	0.195 ± 0.012	0.184 ± 0.008	$\mathbf{0.174 \pm 0.008}$
10	0.181 ± 0.009	0.181 ± 0.009	0.185 ± 0.007	$\mathbf{0.178 \pm 0.009}$	0.188 ± 0.009	0.180 ± 0.007	0.181 ± 0.009
11	0.194 ± 0.008	0.195 ± 0.009	0.189 ± 0.010	0.190 ± 0.008	0.197 ± 0.009	0.188 ± 0.009	$\mathbf{0.183 \pm 0.008}$
12	$\mathbf{0.188 \pm 0.009}$	$\mathbf{0.188 \pm 0.009}$	$\mathbf{0.188 \pm 0.009}$	$\mathbf{0.188 \pm 0.009}$	$\mathbf{0.188 \pm 0.009}$	$\mathbf{0.188 \pm 0.009}$	$\mathbf{0.188 \pm 0.009}$

Table 7. Model pool for MEDMCQA: 12 LLMs with individual accuracies and Gaussian-copula parameters. Results averaged over 6 runs (3 temperatures $\times$ 2 runs). Models sorted by accuracy.

Model	**Accuracy**	**Threshold** τ_j	**Avg.** ρ_j
OpenAI/GPT-5-chat	$83.1\% \pm 0.1\%$	-0.96	0.59
OpenAI/GPT-4.1	$82.2\% \pm 0.1\%$	-0.92	0.53
OpenAI/GPT-4o	$80.4\% \pm 0.3\%$	-0.86	0.65
OpenAI/GPT-4.1-mini	$77.7\% \pm 0.1\%$	-0.76	0.46
Qwen/Qwen3-235b	$77.4\% \pm 0.3\%$	-0.75	0.63
Moonshotai/kimi-k2	$76.3\% \pm 0.6\%$	-0.72	0.63
Google/Gemini-2.5-flash	$72.9\% \pm 0.4\%$	-0.61	0.58
MistralAI/Mistral-3.2-24b	$72.2\% \pm 0.3\%$	-0.59	0.61
Anthropic/Claude-3.5-haiku	$70.7\% \pm 0.2\%$	-0.54	0.52
MistralAI/Mistral-3.1-24b	$69.7\% \pm 0.1\%$	-0.52	0.58
OpenAI/GPT-4.1-nano	$66.8\% \pm 0.3\%$	-0.43	0.51
Meta-LLaMA/LLaMA-3.1-8b	$66.0\% \pm 1.5\%$	-0.41	0.34
Average	74.6%	—	$\bar{\rho} = 0.55$

Table 8. Test error probability on MMLU with MAP aggregation for all methods. Results averaged over 30 evaluations (3 temperatures $\times$ 2 runs $\times$ 5 folds). Bold indicates the best method for each k.

k	Top-k	Term 1	mRMR	Shapley	LogReg	LLM-TOPLA	Greedy MI
1	$\mathbf{0.151 \pm 0.007}$	$\mathbf{0.151 \pm 0.007}$	$\mathbf{0.151 \pm 0.007}$	$\mathbf{0.151 \pm 0.007}$	$\mathbf{0.151 \pm 0.007}$	$\mathbf{0.151 \pm 0.007}$	$\mathbf{0.151 \pm 0.007}$
2	0.152 ± 0.007	0.152 ± 0.007	$\mathbf{0.151 \pm 0.007}$	0.152 ± 0.007	0.152 ± 0.007	0.156 ± 0.008	$\mathbf{0.151 \pm 0.008}$
3	0.152 ± 0.007	0.152 ± 0.008	0.151 ± 0.007	0.152 ± 0.008	0.152 ± 0.008	0.156 ± 0.007	$\mathbf{0.148 \pm 0.008}$
4	0.150 ± 0.005	0.150 ± 0.005	0.150 ± 0.007	0.150 ± 0.005	0.150 ± 0.006	0.150 ± 0.006	$\mathbf{0.144 \pm 0.007}$
5	0.149 ± 0.007	0.149 ± 0.007	0.149 ± 0.008	0.150 ± 0.007	0.150 ± 0.006	0.147 ± 0.006	$\mathbf{0.142 \pm 0.008}$
6	0.149 ± 0.007	0.149 ± 0.007	0.149 ± 0.008	0.149 ± 0.007	0.149 ± 0.007	0.147 ± 0.007	$\mathbf{0.141 \pm 0.007}$
7	0.150 ± 0.007	0.149 ± 0.007	0.151 ± 0.008	0.150 ± 0.007	0.151 ± 0.007	0.148 ± 0.008	$\mathbf{0.143 \pm 0.007}$
8	0.151 ± 0.006	0.151 ± 0.006	0.150 ± 0.007	0.151 ± 0.006	0.152 ± 0.006	0.149 ± 0.006	$\mathbf{0.148 \pm 0.007}$
9	0.153 ± 0.007	0.152 ± 0.007	$\mathbf{0.151 \pm 0.008}$	0.153 ± 0.007	0.155 ± 0.007	$\mathbf{0.151 \pm 0.007}$	$\mathbf{0.151 \pm 0.008}$
10	0.156 ± 0.007	0.156 ± 0.007	0.155 ± 0.008	0.156 ± 0.007	0.155 ± 0.007	$\mathbf{0.152 \pm 0.007}$	0.155 ± 0.007
11	0.158 ± 0.007	0.158 ± 0.007	0.159 ± 0.008	0.158 ± 0.007	0.158 ± 0.007	$\mathbf{0.157 \pm 0.007}$	0.158 ± 0.007
12	$\mathbf{0.164 \pm 0.008}$	$\mathbf{0.164 \pm 0.008}$	0.166 ± 0.007	$\mathbf{0.164 \pm 0.008}$	$\mathbf{0.164 \pm 0.008}$	$\mathbf{0.164 \pm 0.008}$	$\mathbf{0.164 \pm 0.008}$
13	$\mathbf{0.171 \pm 0.008}$	$\mathbf{0.171 \pm 0.008}$	$\mathbf{0.171 \pm 0.008}$	$\mathbf{0.171 \pm 0.008}$	$\mathbf{0.171 \pm 0.008}$	$\mathbf{0.171 \pm 0.008}$	$\mathbf{0.171 \pm 0.008}$

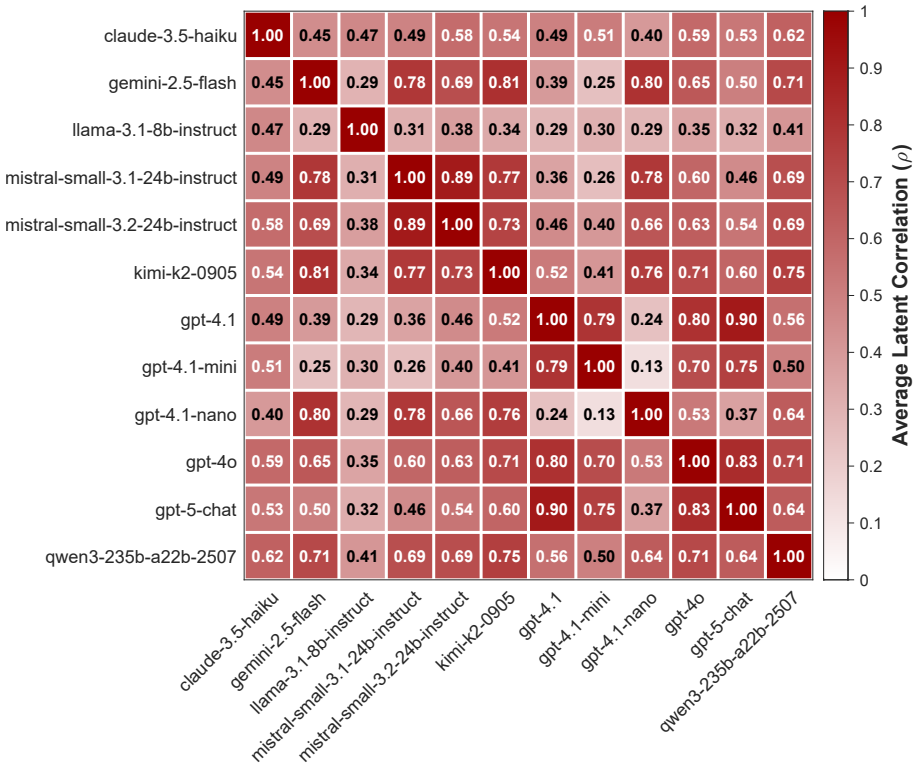


Figure 21. Averaged corr. matrix for MEDMCQA for 3 temperature settings and 2 runs.

Table 9. Test error probability on MMLU with majority voting (MV) aggregation for baselines versus MAP aggregation for Greedy MI. Results averaged over 30 evaluations. Bold indicates the best method for each k.

k	Top-k (MV)	Term 1 (MV)	mRMR (MV)	TMC-Shapley (MV)	LogReg (MV)	LLM-TOPLA (MV)	Greedy MI (MAP)
1	$\mathbf{0.151 \pm 0.007}$	$\mathbf{0.151 \pm 0.007}$	$\mathbf{0.151 \pm 0.007}$	0.156 ± 0.011	$\mathbf{0.151 \pm 0.007}$	$\mathbf{0.151 \pm 0.007}$	$\mathbf{0.151 \pm 0.007}$
2	0.157 ± 0.006	0.157 ± 0.006	0.247 ± 0.022	0.157 ± 0.010	$\mathbf{0.151 \pm 0.008}$	0.183 ± 0.049	$\mathbf{0.151 \pm 0.008}$
3	0.151 ± 0.007	0.151 ± 0.007	0.219 ± 0.026	0.152 ± 0.008	0.158 ± 0.013	0.167 ± 0.011	$\mathbf{0.148 \pm 0.008}$
4	0.151 ± 0.006	0.150 ± 0.007	0.184 ± 0.012	0.150 ± 0.007	0.151 ± 0.009	0.168 ± 0.009	$\mathbf{0.144 \pm 0.007}$
5	0.148 ± 0.006	0.147 ± 0.007	0.201 ± 0.012	0.148 ± 0.008	0.153 ± 0.011	0.162 ± 0.008	$\mathbf{0.141 \pm 0.008}$
6	0.145 ± 0.006	0.145 ± 0.006	0.176 ± 0.011	0.146 ± 0.007	0.151 ± 0.008	0.156 ± 0.007	$\mathbf{0.142 \pm 0.008}$
7	0.148 ± 0.006	0.149 ± 0.006	0.184 ± 0.014	0.149 ± 0.006	0.159 ± 0.012	0.158 ± 0.009	$\mathbf{0.144 \pm 0.006}$
8	$\mathbf{0.149 \pm 0.005}$	$\mathbf{0.149 \pm 0.005}$	0.172 ± 0.010	$\mathbf{0.149 \pm 0.005}$	0.157 ± 0.010	0.155 ± 0.008	$\mathbf{0.149 \pm 0.006}$
9	0.154 ± 0.005	0.154 ± 0.006	0.178 ± 0.008	0.154 ± 0.007	0.166 ± 0.012	0.160 ± 0.007	$\mathbf{0.152 \pm 0.007}$
10	$\mathbf{0.155 \pm 0.006}$	$\mathbf{0.155 \pm 0.005}$	0.167 ± 0.008	$\mathbf{0.155 \pm 0.006}$	0.162 ± 0.009	0.158 ± 0.007	0.157 ± 0.007
11	0.165 ± 0.007	0.165 ± 0.007	0.172 ± 0.009	0.165 ± 0.007	0.170 ± 0.010	0.165 ± 0.008	$\mathbf{0.158 \pm 0.007}$
12	0.164 ± 0.008	0.164 ± 0.008	0.163 ± 0.008	0.164 ± 0.008	0.166 ± 0.007	$\mathbf{0.162 \pm 0.008}$	0.164 ± 0.008
13	0.174 ± 0.008	0.174 ± 0.008	0.174 ± 0.008	0.174 ± 0.008	0.174 ± 0.008	0.174 ± 0.008	$\mathbf{0.171 \pm 0.008}$

Table 10. Test error probability on MMLU with weighted majority voting (WV) aggregation for baselines versus MAP aggregation for Greedy MI. Results averaged over 30 evaluations. Bold indicates the best method for each k.

k	Top-k (WV)	Term 1 (WV)	mRMR (WV)	TMC-Shapley (WV)	LogReg (WV)	LLM-TOPLA (WV)	Greedy MI (MAP)
1	$\mathbf{0.151 \pm 0.007}$	$\mathbf{0.151 \pm 0.007}$	$\mathbf{0.151 \pm 0.007}$	0.154 ± 0.009	$\mathbf{0.151 \pm 0.007}$	$\mathbf{0.151 \pm 0.007}$	$\mathbf{0.151 \pm 0.007}$
2	$\mathbf{0.151 \pm 0.007}$	$\mathbf{0.151 \pm 0.007}$	$\mathbf{0.151 \pm 0.007}$	0.152 ± 0.007	$\mathbf{0.151 \pm 0.007}$	$\mathbf{0.151 \pm 0.007}$	$\mathbf{0.151 \pm 0.008}$
3	0.151 ± 0.007	0.151 ± 0.007	0.219 ± 0.026	0.152 ± 0.008	0.158 ± 0.013	0.167 ± 0.011	$\mathbf{0.148 \pm 0.008}$
4	0.149 ± 0.007	0.149 ± 0.007	0.175 ± 0.009	0.150 ± 0.005	0.151 ± 0.008	0.155 ± 0.009	$\mathbf{0.144 \pm 0.007}$
5	0.148 ± 0.006	0.147 ± 0.007	0.201 ± 0.012	0.148 ± 0.007	0.153 ± 0.011	0.162 ± 0.008	$\mathbf{0.141 \pm 0.008}$
6	0.147 ± 0.006	0.147 ± 0.006	0.170 ± 0.012	0.148 ± 0.006	0.151 ± 0.007	0.155 ± 0.007	$\mathbf{0.142 \pm 0.008}$
7	0.148 ± 0.006	0.149 ± 0.006	0.184 ± 0.014	0.149 ± 0.006	0.159 ± 0.012	0.158 ± 0.009	$\mathbf{0.144 \pm 0.006}$
8	$\mathbf{0.148 \pm 0.006}$	$\mathbf{0.148 \pm 0.006}$	0.169 ± 0.010	0.149 ± 0.006	0.155 ± 0.009	0.154 ± 0.007	0.149 ± 0.006
9	0.154 ± 0.005	0.154 ± 0.006	0.178 ± 0.008	0.155 ± 0.006	0.166 ± 0.012	0.160 ± 0.007	$\mathbf{0.152 \pm 0.007}$
10	$\mathbf{0.153 \pm 0.006}$	0.154 ± 0.006	0.165 ± 0.008	$\mathbf{0.153 \pm 0.006}$	0.160 ± 0.008	0.154 ± 0.006	0.157 ± 0.007
11	0.165 ± 0.007	0.165 ± 0.007	0.171 ± 0.009	0.165 ± 0.007	0.170 ± 0.010	0.165 ± 0.008	$\mathbf{0.158 \pm 0.007}$
12	0.163 ± 0.007	0.163 ± 0.007	0.162 ± 0.008	0.162 ± 0.007	0.164 ± 0.007	$\mathbf{0.160 \pm 0.008}$	0.164 ± 0.008
13	0.174 ± 0.009	0.174 ± 0.009	0.174 ± 0.009	0.174 ± 0.009	0.174 ± 0.009	0.174 ± 0.009	$\mathbf{0.171 \pm 0.008}$

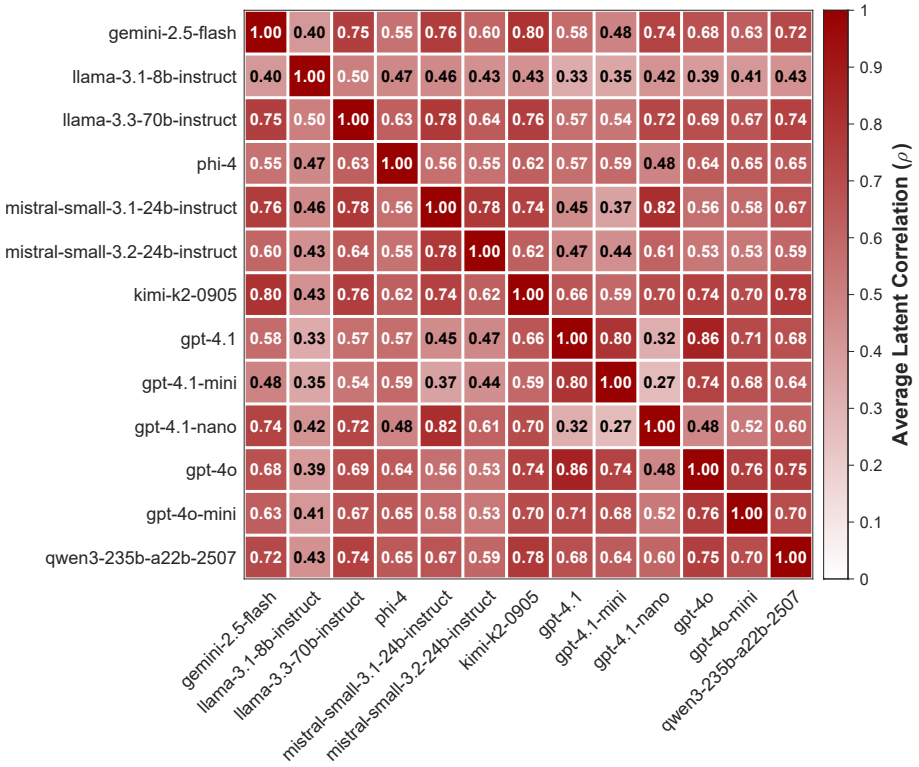


Figure 22. Averaged corr. matrix for MMLU for 3 temperature settings and 2 runs.

Table 11. Model pool for MMLU: 13 LLMs with individual accuracies and Gaussian-copula parameters. Results averaged over 6 runs (3 temperatures $\times$ 2 runs). Models sorted by accuracy.

Model	**Accuracy**	**Threshold** τ_j	**Avg.** ρ_j
OpenAI/GPT-4.1	$85.0\% \pm 0.2\%$	-1.04	0.58
OpenAI/GPT-4o	$84.0\% \pm 0.3\%$	-0.99	0.65
Qwen/Qwen3-235b	$82.2\% \pm 0.2\%$	-0.92	0.66
Moonshotai/kimi-k2	$82.1\% \pm 0.8\%$	-0.92	0.68
OpenAI/GPT-4.1-mini	$81.6\% \pm 0.2\%$	-0.90	0.54
Google/Gemini-2.5-flash	$79.8\% \pm 0.4\%$	-0.83	0.64
OpenAI/GPT-4o-mini	$78.3\% \pm 0.2\%$	-0.78	0.63
Meta-LLaMA/LLaMA-3.3-70b	$77.2\% \pm 0.1\%$	-0.75	0.67
microsoft/phi-4	$74.1\% \pm 0.5\%$	-0.65	0.58
MistralAI/Mistral-3.2-24b	$73.2\% \pm 0.5\%$	-0.62	0.57
MistralAI/Mistral-3.1-24b	$71.8\% \pm 0.1\%$	-0.58	0.63
OpenAI/GPT-4.1-nano	$69.2\% \pm 0.2\%$	-0.50	0.56
Meta-LLaMA/LLaMA-3.1-8b	$65.9\% \pm 1.3\%$	-0.41	0.42
Average	77.3%	—	$\bar{\rho} = 0.60$

Table 12. Test error probability on IMDB movie reviews with MAP aggregation for all methods. Results averaged over 30 evaluations (3 temperatures $\times$ 2 runs $\times$ 5 splits). Bold indicates the best method for each k.

k	Top-k	Term 1	mRMR	TMC-Shapley	LogReg	LLM-TOPLA	Greedy MI
1	$\mathbf{0.035 \pm 0.004}$	$\mathbf{0.035 \pm 0.004}$	$\mathbf{0.035 \pm 0.004}$	0.038 ± 0.005	$\mathbf{0.035 \pm 0.004}$	$\mathbf{0.035 \pm 0.004}$	$\mathbf{0.035 \pm 0.004}$
2	0.036 ± 0.004	0.036 ± 0.004	$\mathbf{0.035 \pm 0.004}$	0.036 ± 0.004	$\mathbf{0.035 \pm 0.004}$	$\mathbf{0.035 \pm 0.004}$	$\mathbf{0.035 \pm 0.004}$
3	$\mathbf{0.034 \pm 0.004}$	$\mathbf{0.034 \pm 0.004}$	0.035 ± 0.004	0.035 ± 0.004	0.035 ± 0.004	0.035 ± 0.004	0.036 ± 0.004
4	0.036 ± 0.004	0.036 ± 0.004	$\mathbf{0.034 \pm 0.004}$	0.035 ± 0.004	0.035 ± 0.004	0.035 ± 0.004	$\mathbf{0.034 \pm 0.004}$
5	0.038 ± 0.004	0.038 ± 0.004	0.034 ± 0.004	0.036 ± 0.004	0.035 ± 0.004	0.034 ± 0.004	$\mathbf{0.033 \pm 0.004}$
6	0.039 ± 0.004	0.039 ± 0.004	0.034 ± 0.004	0.037 ± 0.003	0.036 ± 0.004	0.035 ± 0.004	$\mathbf{0.033 \pm 0.004}$
7	0.040 ± 0.004	0.039 ± 0.004	0.035 ± 0.004	0.038 ± 0.004	0.036 ± 0.004	0.035 ± 0.004	$\mathbf{0.034 \pm 0.004}$
8	0.040 ± 0.004	0.040 ± 0.004	0.036 ± 0.005	0.038 ± 0.004	0.037 ± 0.004	0.036 ± 0.004	$\mathbf{0.035 \pm 0.004}$
9	0.039 ± 0.004	0.039 ± 0.004	0.038 ± 0.004	0.039 ± 0.004	0.038 ± 0.004	0.038 ± 0.004	$\mathbf{0.037 \pm 0.004}$
10	0.040 ± 0.004	0.040 ± 0.004	0.039 ± 0.004	0.040 ± 0.004	0.039 ± 0.004	$\mathbf{0.038 \pm 0.004}$	$\mathbf{0.038 \pm 0.004}$
11	0.041 ± 0.004	0.041 ± 0.004	$\mathbf{0.040 \pm 0.004}$	0.041 ± 0.005	$\mathbf{0.040 \pm 0.004}$	$\mathbf{0.040 \pm 0.004}$	$\mathbf{0.040 \pm 0.004}$
12	$\mathbf{0.041 \pm 0.005}$	$\mathbf{0.041 \pm 0.005}$	$\mathbf{0.041 \pm 0.005}$	$\mathbf{0.041 \pm 0.005}$	$\mathbf{0.041 \pm 0.005}$	$\mathbf{0.041 \pm 0.005}$	$\mathbf{0.041 \pm 0.005}$

Table 13. Test error probability on IMDB movie reviews with majority voting (MV) aggregation for baselines versus MAP aggregation for Greedy MI. Results averaged over 30 evaluations. Bold indicates the best method for each k.

k	Top-k (MV)	Term 1 (MV)	mRMR (MV)	TMC-Shapley (MV)	LogReg (MV)	LLM-TOPLA (MV)	Greedy MI (MAP)
1	$\mathbf{0.035 \pm 0.004}$	$\mathbf{0.035 \pm 0.004}$	$\mathbf{0.035 \pm 0.004}$	0.039 ± 0.006	$\mathbf{0.035 \pm 0.004}$	$\mathbf{0.035 \pm 0.004}$	$\mathbf{0.035 \pm 0.004}$
2	$\mathbf{0.035 \pm 0.004}$	$\mathbf{0.035 \pm 0.004}$	$\mathbf{0.035 \pm 0.004}$	0.038 ± 0.005	0.038 ± 0.005	0.036 ± 0.004	$\mathbf{0.035 \pm 0.004}$
3	$\mathbf{0.034 \pm 0.004}$	$\mathbf{0.034 \pm 0.004}$	0.039 ± 0.004	0.036 ± 0.004	0.036 ± 0.005	0.041 ± 0.004	0.036 ± 0.004
4	0.036 ± 0.004	0.036 ± 0.004	0.037 ± 0.004	0.037 ± 0.004	0.036 ± 0.004	0.037 ± 0.004	$\mathbf{0.034 \pm 0.004}$
5	0.036 ± 0.004	0.036 ± 0.004	0.038 ± 0.004	0.036 ± 0.004	0.038 ± 0.005	0.040 ± 0.004	$\mathbf{0.033 \pm 0.004}$
6	0.036 ± 0.004	0.036 ± 0.004	0.036 ± 0.004	0.037 ± 0.004	0.038 ± 0.004	0.037 ± 0.004	$\mathbf{0.033 \pm 0.004}$
7	0.036 ± 0.004	0.036 ± 0.004	0.038 ± 0.004	0.037 ± 0.004	0.040 ± 0.005	0.040 ± 0.004	$\mathbf{0.034 \pm 0.004}$
8	0.036 ± 0.004	0.036 ± 0.004	0.037 ± 0.004	0.037 ± 0.004	0.039 ± 0.004	0.038 ± 0.004	$\mathbf{0.035 \pm 0.004}$
9	$\mathbf{0.037 \pm 0.004}$	$\mathbf{0.037 \pm 0.004}$	0.039 ± 0.004	0.038 ± 0.004	0.040 ± 0.004	0.040 ± 0.005	$\mathbf{0.037 \pm 0.004}$
10	0.038 ± 0.004	0.038 ± 0.004	$\mathbf{0.037 \pm 0.004}$	0.038 ± 0.004	0.039 ± 0.004	0.039 ± 0.004	0.038 ± 0.004
11	$\mathbf{0.040 \pm 0.004}$	$\mathbf{0.040 \pm 0.004}$	$\mathbf{0.040 \pm 0.004}$	$\mathbf{0.040 \pm 0.004}$	0.041 ± 0.004	0.041 ± 0.004	$\mathbf{0.040 \pm 0.004}$
12	$\mathbf{0.040 \pm 0.004}$	$\mathbf{0.040 \pm 0.004}$	$\mathbf{0.040 \pm 0.004}$	$\mathbf{0.040 \pm 0.004}$	$\mathbf{0.040 \pm 0.004}$	$\mathbf{0.040 \pm 0.004}$	0.041 ± 0.005

Table 14. Test error probability on IMDB movie reviews with weighted majority voting (WV) aggregation for baselines versus MAP aggregation for Greedy MI. Results averaged over 30 evaluations. Bold indicates the best method for each k.

k	Top-k (WV)	Term 1 (WV)	mRMR (WV)	TMC-Shapley (WV)	LogReg (WV)	LLM-TOPLA (WV)	Greedy MI (MAP)
1	$\mathbf{0.035 \pm 0.004}$	$\mathbf{0.035 \pm 0.004}$	$\mathbf{0.035 \pm 0.004}$	0.039 ± 0.005	$\mathbf{0.035 \pm 0.004}$	$\mathbf{0.035 \pm 0.004}$	$\mathbf{0.035 \pm 0.004}$
2	$\mathbf{0.035 \pm 0.004}$	$\mathbf{0.035 \pm 0.004}$	$\mathbf{0.035 \pm 0.004}$	0.036 ± 0.004	$\mathbf{0.035 \pm 0.004}$	$\mathbf{0.035 \pm 0.004}$	$\mathbf{0.035 \pm 0.004}$
3	$\mathbf{0.034 \pm 0.004}$	$\mathbf{0.034 \pm 0.004}$	0.039 ± 0.004	0.036 ± 0.004	0.036 ± 0.005	0.041 ± 0.004	0.036 ± 0.004
4	$\mathbf{0.034 \pm 0.004}$	$\mathbf{0.034 \pm 0.004}$	0.035 ± 0.004	0.035 ± 0.004	0.036 ± 0.004	0.037 ± 0.004	$\mathbf{0.034 \pm 0.004}$
5	0.036 ± 0.004	0.036 ± 0.004	0.038 ± 0.004	0.036 ± 0.004	0.038 ± 0.005	0.040 ± 0.004	$\mathbf{0.033 \pm 0.004}$
6	0.035 ± 0.004	0.035 ± 0.004	0.036 ± 0.004	0.035 ± 0.004	0.037 ± 0.004	0.037 ± 0.004	$\mathbf{0.033 \pm 0.004}$
7	0.036 ± 0.004	0.036 ± 0.004	0.038 ± 0.004	0.037 ± 0.005	0.040 ± 0.005	0.040 ± 0.004	$\mathbf{0.034 \pm 0.004}$
8	$\mathbf{0.035 \pm 0.004}$	$\mathbf{0.035 \pm 0.004}$	0.037 ± 0.004	0.036 ± 0.004	0.039 ± 0.004	0.038 ± 0.004	$\mathbf{0.035 \pm 0.004}$
9	$\mathbf{0.037 \pm 0.004}$	$\mathbf{0.037 \pm 0.004}$	0.039 ± 0.004	0.038 ± 0.004	0.040 ± 0.004	0.040 ± 0.005	$\mathbf{0.037 \pm 0.004}$
10	$\mathbf{0.037 \pm 0.004}$	$\mathbf{0.037 \pm 0.004}$	$\mathbf{0.037 \pm 0.004}$	$\mathbf{0.037 \pm 0.004}$	0.039 ± 0.004	0.039 ± 0.004	0.038 ± 0.004
11	$\mathbf{0.040 \pm 0.004}$	$\mathbf{0.040 \pm 0.004}$	$\mathbf{0.040 \pm 0.004}$	$\mathbf{0.040 \pm 0.004}$	0.041 ± 0.004	0.041 ± 0.004	$\mathbf{0.040 \pm 0.004}$
12	$\mathbf{0.040 \pm 0.004}$	$\mathbf{0.040 \pm 0.004}$	$\mathbf{0.040 \pm 0.004}$	$\mathbf{0.040 \pm 0.004}$	$\mathbf{0.040 \pm 0.004}$	$\mathbf{0.040 \pm 0.004}$	0.041 ± 0.005

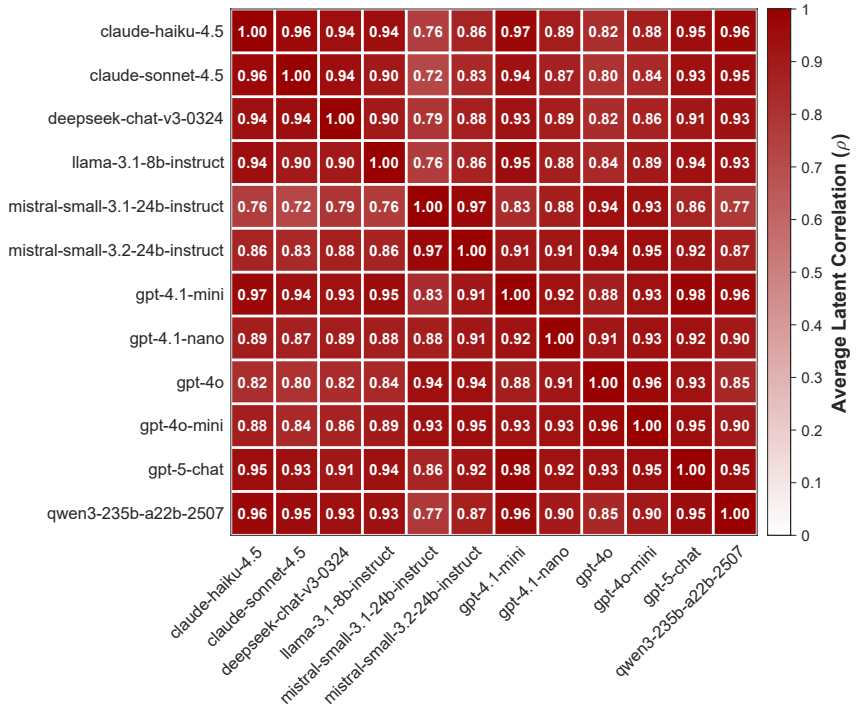


Figure 23. Averaged corr. matrix for IMDB movie reviews for 3 temperature settings and 2 runs.

Table 15. Model pool for IMDB movie reviews dataset: 12 LLMs with individual accuracies and Gaussian-copula parameters. Results averaged over 6 runs (3 temperatures $\times$ 2 runs). Models sorted by accuracy.

Model	**Accuracy**	**Threshold** τ_j	**Avg.** ρ_j
Anthropic/Claude-sonnet-4.5	$96.39\% \pm 0.05\%$	-1.798	0.881
OpenAI/GPT-5-chat	$96.26\% \pm 0.07\%$	-1.782	0.930
Anthropic/Claude-haiku-4.5	$96.10\% \pm 0.08\%$	-1.762	0.902
OpenAI/GPT-4.1-mini	$95.85\% \pm 0.08\%$	-1.733	0.926
Qwen/Qwen3-235b	$95.67\% \pm 0.09\%$	-1.714	0.907
deepseek/deepseek-chat-v3	$95.37\% \pm 0.10\%$	-1.682	0.891
Meta-LLaMA/LLaMA-3.1-8b	$94.79\% \pm 0.17\%$	-1.625	0.890
OpenAI/GPT-4o	$94.49\% \pm 0.05\%$	-1.597	0.880
OpenAI/GPT-4.1-nano	$94.19\% \pm 0.07\%$	-1.571	0.900
OpenAI/GPT-4o-mini	$94.00\% \pm 0.07\%$	-1.555	0.909
MistralAI/Mistral-small-3.2-24b	$93.60\% \pm 0.18\%$	-1.522	0.900
MistralAI/Mistral-small-3.1-24b	$91.81\% \pm 0.06\%$	-1.393	0.84
Average	94.54%	—	$\bar{\rho} = 0.90$