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

Yigit Turkmen
Dept. of Electrical and Electronics Eng.
Bilkent University
Ankara, Turkey
yigit.turkmen@ug.bilkent.edu.tr

Baturalp Buyukates
School of Computer Science
University of Birmingham
Birmingham, UK
b.buyukates@bham.ac.uk

Melih Bastopcu
Dept. of Electrical and Electronics Eng.
Bilkent University
Ankara, Turkey
bastopcu@bilkent.edu.tr

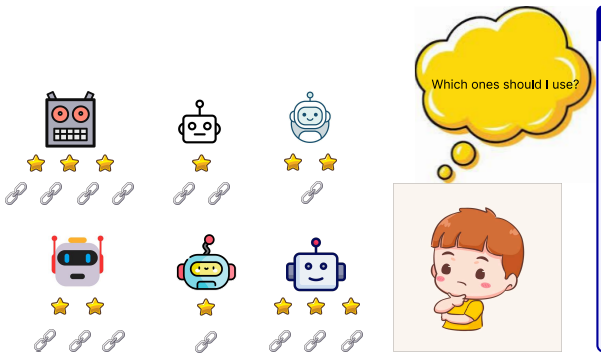
Abstract

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 a Gaussian copula and show that the error of Naive Bayes (plurality) aggregation converges to a strictly positive floor as the ensemble grows. 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 on four datasets spanning binary and multi-class classification and sentiment analysis: MEDMCQA, MMLU, AG News, and IMDB. Compared with eleven baselines under the same query budget, our method attains the lowest or tied-lowest test error for ensembles of three to seven models on MEDMCQA, MMLU, and AG News, while all methods perform comparably on IMDB. The source code, experiment configurations, and instructions for accessing the datasets and reproducing the results are available in the project repository.

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 monetary 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 multi-class (M-ary) classification under an arbitrary class prior.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).



(a) Candidate pool.

AG News: Correlated Failure

Text: China's inflation rate slows sharply but problems remain (AFP) - China's inflation rate eased sharply in October as government efforts to cool the economy began to really bite...
Ground Truth: `WORLD`

Model	Acc.	Ans.
Qwen3-235B	88%	`BUSINESS` ✗
Gemini-2.5-F.	84%	`BUSINESS` ✗
GPT-4.1-mini	82%	`BUSINESS` ✗
Llama-3.1-8B	76%	`WORLD` ✓
Avg.	**82.5%**	**1/4** ✓

(b) Correlated failure.

Figure 1: **(a)** Candidate LLM pool for ensembling: stars indicate model accuracy, while chain links denote pairwise correlation (shared error patterns) with other models. **(b)** An AG News sample where three high-accuracy models each misclassify a macroeconomic text as `BUSINESS`. The smaller Llama-3.1-8B model accurately predicts `WORLD`, demonstrating the critical value of ensemble diversity.

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 $\mathcal{S}$ of size k that minimizes the final decision error under an optimal (or near-optimal) aggregator as shown in Figure 1a. 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 1b. 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, treating each model output as a noisy observation of the label and using mutual information to quantify subset informativeness. Our analysis covers both the independent and Gaussian-copula correlated regimes, and motivates a greedy selection algorithm.

Contributions:

- We model LLM responses with a Gaussian-copula capturing both individual accuracies and inter-model correlations, and validate it empirically on real LLM ensembles.
- We prove that under independent errors, Top-k by accuracy is optimal (Theorem 3.1). Under correlated errors, we derive a decomposition (Theorem 3.2) that augments mRMR Peng et al. (2005) with an error-correlation term, based on which we propose a greedy algorithm.
- We characterize a non-vanishing error floor for Naive Bayes (plurality) aggregation of equicorrelated ensembles (Theorem 3.3), explaining why scaling correlated ensembles cannot eliminate the error of aggregators that ignore the dependence between models.
- We evaluate against eleven baselines under identical query budgets on MEDMCQA, MMLU, AG News, and IMDB (binary and multi-class): Greedy MI attains the lowest or tied-lowest error at mid-range budgets on the first three, and all methods perform comparably on IMDB.

A comprehensive review of related work is provided in Appendix A.

2 System Model and Problem Formulation

In this work, we consider a multi-class (M-ary) classification problem carried out by a group of LLMs, where each incoming task involves selecting one label from $\mathcal{M} \triangleq \{1, \ldots, M\}$. We model the correct label with a random variable $Y \in \mathcal{M}$ 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 a pool of K LLMs, indexed by the set $\mathcal{K} \triangleq \{1, \ldots, K\}$. For a given input instance, model j produces an M-ary prediction $X_j \in \mathcal{M}$.

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

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

The error indicator for LLM j is $E_j \triangleq 1(X_j \neq Y)$, with marginal error probability $\mathbb{P}(E_j = 1) = \epsilon_j$ and accuracy $\alpha_j = 1 - \epsilon_j$. While our baseline analysis (Section 3.1) evaluates the special regime where the predictions X_j are strictly conditionally independent given Y, real-world LLMs rarely operate in this regime. In practice, models share training data and architectures, necessitating a framework that captures these inter-model dependencies. We denote the vector of responses for a subset $\mathcal{S} \subseteq \mathcal{K}$ as $\mathbf{X}_\mathcal{S} \triangleq (X_j)_{j \in \mathcal{S}}$.

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, 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. This construction decouples marginal accuracies from the dependence structure: τ_j controls ϵ_j, while Σ controls error correlations. Crucially, while the Gaussian-copula captures the correlation of the binary error events (i.e., whether an error occurs simultaneously), we maintain the M-ary Symmetric Channel assumption that, conditional on an error occurring ($E_j = 1$), the specific incorrect label is drawn uniformly and independently from the remaining $M-1$ classes. This specifies the full joint distribution of the ensemble responses required for our theoretical analysis.

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 $\mathcal{S}$ and their observed outputs $\mathbf{X}_\mathcal{S} = \mathbf{x}_\mathcal{S}$, the MAP rule is given by

$$\widehat{Y}_{\mathrm{MAP}}(\mathbf{x}_\mathcal{S}) = \operatorname*{argmax}_{y \in \mathcal{M}} \pi_y\, \mathbb{P}(\mathbf{X}_\mathcal{S} = \mathbf{x}_\mathcal{S} \mid Y = y). \tag{4}$$

The performance of any subset $\mathcal{S}$ is measured by its error probability given by $P_e(\mathcal{S}) \triangleq \mathbb{P}(\widehat{Y}_{\mathrm{MAP}}(\mathbf{X}_\mathcal{S}) \neq Y)$.

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 $\mathcal{S}_k^\star$ of size k minimizing the probability of error:

$$\mathcal{S}_k^\star \in \operatorname*{arg\,min}_{\mathcal{S} \subseteq \mathcal{K},\, |\mathcal{S}| = k} P_e(\mathcal{S}). \tag{5}$$

Finding $\mathcal{S}_k^\star$ 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 3.3.

3 Information-Theoretic Analysis of Ensemble Selection

The central challenge in ensemble selection is that the error-optimal subset $\mathcal{S}^\star$ 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 show that accuracy-based and information-theoretic selection coincide under independent errors (Section 3.1), then decompose how correlated errors change the marginal value of a model (Section 3.2), and finally derive the error floor of Naive Bayes aggregation in correlated ensembles (Section 3.3).

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) Proakis & Salehi (2008), 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 $\mathcal{M}$. 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 $\mathcal{S}_k^{MI} \in \operatorname{argmax}_{|\mathcal{S}|=k} I(Y; \mathbf{X}_\mathcal{S})$ denote the subset of size k that maximizes the mutual information between Y and $\mathbf{X}_\mathcal{S}$. For any subset $\mathcal{S} \subseteq \mathcal{K}$ with $|\mathcal{S}| = k$, the top-k set H_k dominates $\mathcal{S}$ both in terms of maximizing the mutual information and minimizing the probability of error. Thus, we have*

$$I(Y; \mathbf{X}_{\mathbf{H_k}}) \geq I(Y; \mathbf{X}_\mathcal{S}), \tag{6}$$

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

Consequently, when the channels are independent and symmetric, H_k is both an error-optimal subset and a mutual-information-optimal subset of size k, regardless of the label cardinality M and the class prior $\boldsymbol{\pi}$.

The proof of Theorem 3.1 is provided in Appendix B.

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 3.2 (Accuracy-Redundancy-Error Decomposition)**.** *Assuming that ensemble outputs $\mathbf{X}_\mathcal{K}$ follow the M-ary Symmetric Channel (MSC) model, such that conditional on an error occurring ($E_j = 1$), the incorrect prediction is drawn uniformly from the remaining $M-1$ classes. Then, for LLM outputs $X_j \in \mathcal{M}$ and true label $Y \in \mathcal{M}$ 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 $\mathcal{S}$ as*

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

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

$$\Delta(j \,|\, \mathcal{S}) = I(Y; X_j) - I(X_j; \mathbf{X}_\mathcal{S}) + I(E_j; E_\mathcal{S}) + \Lambda_j(\mathcal{S}), \tag{9}$$

where $E_{\mathcal{S}} = (E_i)_{i\in\mathcal{S}}$, and the last term in (9) *is the label-dependence correction term given by*

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

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

$$\Delta(j \mid \mathcal{S}) = I(Y; X_j) - I(X_j; \mathbf{X}_{\mathcal{S}}) + I(E_j; E_{\mathcal{S}}). \tag{11}$$

The proof of Theorem 3.2 is provided in Appendix C. Under the label-invariant assumption, the decomposition has an intuitive reading: $I(Y; X_j)$ rewards relevance, $I(X_j; \mathbf{X}_{\mathcal{S}})$ penalizes prediction redundancy, and $I(E_j; E_{\mathcal{S}})$—counter-intuitively—rewards predictable error structure (anti-correlated errors are a special case), since knowing the error pattern of $\mathcal{S}$ helps decode Y. Standard mRMR-style criteria (Peng et al., 2005) keep only the first two terms. The third term is the class-conditional redundancy $I(X_j; \mathbf{X}_{\mathcal{S}} \mid Y)$, which appears in information-theoretic expansions of ensemble and feature-set mutual information (Brown, 2009; Brown et al., 2012). Theorem 3.2 shows that, under the MSC model, it equals the error dependence $I(E_j; E_{\mathcal{S}})$ plus the correction $\Lambda_j(\mathcal{S})$.

3.3 Asymptotic Saturation Limit

We now ask: *can ensemble error vanish as $K \to \infty$ under correlated errors?* For equicorrelated Gaussian-copula errors, we show that it cannot when the outputs are aggregated as if they were conditionally independent: as $K \to \infty$, the error probability of Naive Bayes aggregation converges to a strictly positive floor.

Theorem 3.3 (Error Floor under NB Aggregation)**.** *Suppose the model errors follow the equicorrelated Gaussian-copula model of Section 2.1 with correlation $\rho \in (0, 1)$, every model has the same accuracy $\alpha \triangleq 1 - \epsilon > 1/M$, and whenever $E_j = 1$, model j outputs one of the $M - 1$ incorrect labels uniformly at random, independently of the other models. Let the prior be balanced, $\boldsymbol{\pi} = (1/M, \ldots, 1/M)$, and let $\widehat{Y}_{\mathrm{NB}}$ be the Naive Bayes aggregator, which treats the outputs $\mathbf{X}_{\mathcal{K}}$ as conditionally independent given Y and sums their log-likelihoods (here, this reduces to plurality voting). Then, we have*

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

The proof of Theorem 3.3 is given in Appendix D. The theorem shows that the error floor is strict for aggregation rules that treat the outputs as conditionally independent given Y, such as Naive Bayes, which here reduces to plurality voting. Because these rules ignore the common latent factor U, adding more models cannot remove the shared errors; under such rules, gains must instead come from combining models whose errors are less correlated, which is what subset selection does. The floor is therefore a property of the aggregator as well as of the correlation. Under the MSC model with $M \geq 3$, the wrong votes in a large ensemble split evenly across the incorrect labels, so the true label is the only one whose vote share differs from the rest, and an aggregator that uses the joint distribution of the outputs can exploit this. In real LLM ensembles, however, models that are wrong usually choose the same wrong label, which removes this signal. Corollary E.1 in Appendix E shows that the Naive Bayes limit is unchanged under imbalanced priors.

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

Motivated by Theorem 3.2, we propose a greedy strategy that approximates the optimal subset $\mathcal{S}^\star$ by stepwise maximization, since solving (5) exactly requires a search over all $\binom{K}{k}$ subsets. Under a latent-difficulty assumption, the difficulty-aware part of the objective is monotone submodular, so greedy maximization of that part has a $(1 - 1/e)$ guarantee (Theorem F.1, Appendix F) where this guarantee does not cover the full objective. We construct the subset $\mathcal{S}$ iteratively. Let $\mathcal{S}_{k-1}$ 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

$$j^\star = \operatorname*{argmax}_{j\in\mathcal{K}\setminus\mathcal{S}_{k-1}} \Delta(j \mid \mathcal{S}_{k-1}). \tag{13}$$

Algorithm 1 Greedy MI Selection

Input: Estimate the dist. of set outputs $X_{\mathcal{K}}$, labels Y, target size k
Initialize $\mathcal{S} \leftarrow \emptyset$
for $t = 1$ **to** k **do**
  Estimate the full conditional mutual information score $\hat{\Delta}(j|\mathcal{S}) = \widehat{I}(Y; X_j \mid \mathbf{X}_{\mathcal{S}})$ directly using smoothed empirical distributions (see Algorithm 2).
  $j^\star \leftarrow \operatorname{argmax}_j \hat{\Delta}(j|\mathcal{S})$
  $\mathcal{S} \leftarrow \mathcal{S} \cup \{j^\star\}$
end for
Output: Selected subset $\mathcal{S}_k^* = \mathcal{S}$

While Theorem 3.2 provides the theoretical motivation for our approach by decomposing the information gain into accuracy, redundancy, and error correlation, calculating these separated terms explicitly requires the assumption that the label-dependence correction $\Lambda_j(\mathcal{S})$ is negligible. To avoid relying on this theoretical simplification and to ensure that any label-dependent error structures present in real-world LLMs are fully captured, our practical implementation bypasses the decomposition. Instead, as detailed in Algorithm 1, we directly evaluate candidates by estimating the full conditional mutual information $\Delta(j \mid \mathcal{S}) = \widehat{I}(Y; X_j \mid \mathbf{X}_{\mathcal{S}})$ using empirical counts from the training split of each evaluation (Section 5). This split is used only to estimate these statistics and to fit the empirical MAP aggregator; no LLM parameters are trained. The procedure is the same for any number of classes M, and we apply it to both binary (MEDMCQA, MMLU, IMDB) and 4-class (AG News) tasks.

5 Experimental Results

5.1 Experiments on MEDMCQA and MMLU

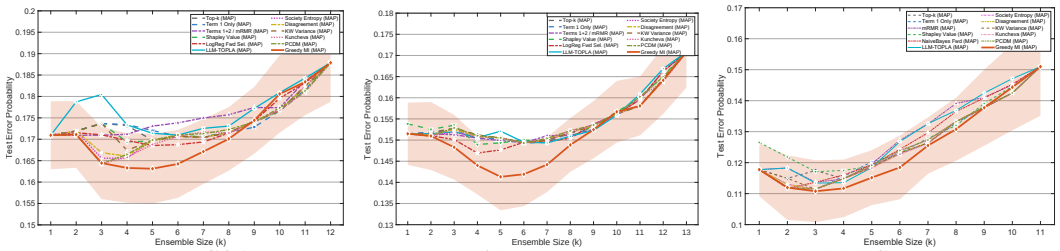
We first evaluate on two constructed binary QA benchmarks. Binary classification captures the canonical LLM-as-judge deployment shape – verification, fact-checking, sentiment, safety, claim entailment, and reward modeling – and admits clean copula validation via tetrachoric correlation, making it a natural starting point Inan et al. (2023); Zheng et al. (2023); Gu et al. (2025); Thorne et al. (2018); Laban et al. (2022); Zhang et al. (2024); Ouyang et al. (2022). We use MEDMCQA Pal et al. (2022), a medical QA dataset whose 4183-sample validation set induces shared failure modes across LLMs, and MMLU Hendrycks et al. (2021), from which we sample 5000 questions across its diverse 57 subjects.

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 for MEDMCQA and 10000 for MMLU.

Evaluation protocol. In each evaluation, every method selects its subset and fits the MAP aggregator on a random 80% training split, and we measure test error on the held-out 20%. For MEDMCQA and MMLU, we repeat this at three sampling temperatures (0.01, 0.3, 0.7), with two independent runs per temperature and five random splits per run, and report the mean test error over all $3 \times 2 \times 5 = 30$ evaluations. Temperatures, runs, and splits serve only to measure robustness: no method observes the temperature, and selection uses only the training split of the current evaluation.

Model pools. The MEDMCQA pool comprises 12 LLMs: GPT-4.1, GPT-4.1-mini, GPT-4.1-nano, GPT-4o, and GPT-5-chat OpenAI (2026); LLaMA-3.1-8b Grattafiori et al. (2024); Qwen3-235b Yang et al. (2025a); kimi-k2 Team et al. (2025); Claude-3.5-haiku Anthropic (2024); Mistral-small-3.1-24b and Mistral-small-3.2-24b Mistral AI (2026); and Gemini-2.5-flash Comanici et al. (2025). Individual accuracies span 66–83% (mean accuracy $\bar{\alpha} = 74.6\%$), with average pairwise correlation $\bar{\rho} = 0.55$. The MMLU pool replaces GPT-5-chat and Claude-3.5-haiku with GPT-4o-mini, LLaMA-3.3-70b and phi-4 Abdin et al. (2024), yielding 13 LLMs with accuracies 66–85% ($\bar{\alpha} = 77.3\%$) and $\bar{\rho} = 0.60$. For all datasets, per-model accuracies and full correlation matrices are in Appendix L.

Baselines. We compare Greedy MI with eleven selection strategies on every dataset. *Top-k* selects the k most accurate models, and *Term 1* ranks models by $\hat{I}(Y; X_j)$, the first term of (11). *mRMR* (Peng et al., 2005) uses the first two terms of (11) but omits the error-dependence term. *TMC-Shapley* (Ghorbani & Zou, 2019) ranks models by Shapley values estimated from 5000 Monte



(a) MEDMCQA (b) MMLU (c) AG News

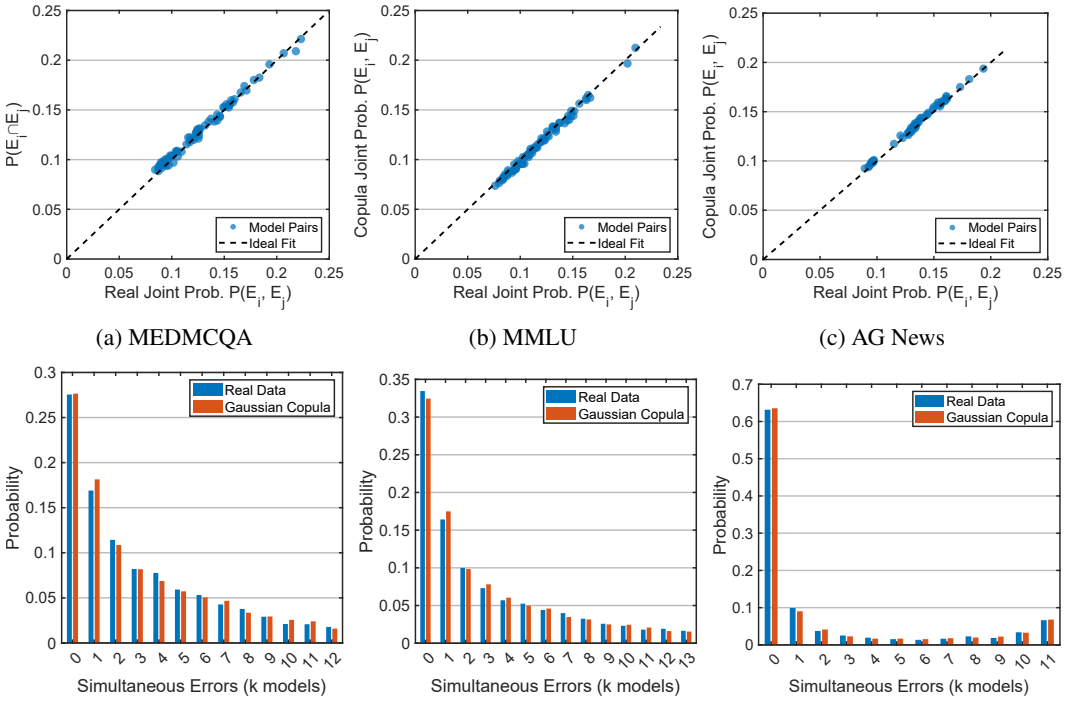
Figure 2: Test error vs. ensemble size on (a) MEDMCQA, (b) MMLU, and (c) AG News, averaged over temperatures, runs, and random splits. All methods use MAP aggregation. The shaded region shows one standard deviation of Greedy MI's error across the 30 evaluations.

Carlo permutations, with the MAP training error as the utility. *Forward Selection* greedily adds the model that most improves 5-fold cross-validated accuracy (logistic regression on the binary datasets, naive Bayes on AG News). *LLM-TOPLA* (Tekin et al., 2024) searches all subsets of size k for the best trade-off between focal diversity and ensemble accuracy. We also include five diversity-based baselines: Society Entropy (Cunningham & Carney, 2000), Disagreement, Kohavi–Wolpert variance, and Kuncheva diversity (Kuncheva & Whitaker, 2003), as well as PCDM (Banfield et al., 2005). All methods use the MAP aggregator at evaluation time. Implementation details are provided in Appendix G for all datasets and algorithms.

Results. Figures 2a and 2b show test error versus ensemble size. Greedy MI achieves the lowest or tied-lowest mean test error in the mid-budget regime $k \in [3, 7]$ on both datasets, including the five diversity-based baselines. At $k = 5$, it achieves 16.3% on MEDMCQA versus 16.9% for the strongest baselines (LogReg and Kuncheva), and 14.1% on MMLU versus 14.8% for LogReg. These gains also allow smaller ensembles: with only $k = 4$ models on MMLU (14.4%), Greedy MI outperforms every baseline at any budget, while its $k = 3$ result on MEDMCQA (16.4%) matches the best baseline result, attained by Kuncheva and PCDM. Ranking by accuracy alone can even hurt: on MEDMCQA, Top-k error rises from 17.1% at $k = 1$ to 17.4% at $k = 3$. Top-k and Term 1 perform almost identically, consistent with the binary symmetric-error setting with balanced labels, where $I(Y; X_j)$ increases monotonically with individual accuracy above chance. Their similar performance highlights the limited benefit of individual relevance alone compared with accounting for joint dependence. At large k, the curves converge because the selected subsets increasingly overlap, and error rises across methods, consistent with the MAP aggregator having to estimate 2^k prediction-pattern probabilities from finite training data.

Copula validation. Figures 3a, b, d, e show two diagnostics for MEDMCQA and MMLU at temp $= 0.7$, run 1. The pairwise scatters (top row) confirm second-order fit by construction: empirical joint error probabilities $P(E_i \cap E_j)$ cluster tightly along the diagonal across all $\binom{12}{2} = 66$ pairs on MEDMCQA and $\binom{13}{2} = 78$ pairs on MMLU. The simultaneous-error histograms (bottom row) confirm higher-order fit on quantities not used during fitting: the copula-predicted distribution (orange) closely matches the observed tail (blue) on both datasets. The same pattern holds across all temperature/run pairs (Appendix H.3, I.3), supporting the Gaussian-copula as an appropriate model for LLM error dependence.

Selection behavior. The methods select very different ensembles. mRMR's redundancy penalty pulls in low-accuracy models early: LLaMA-3.1-8b (66.0% accuracy) at $k = 2$ on MEDMCQA, and LLaMA-3.3-70b (77.2%) instead of Qwen3-235b (82.2%) at $k = 2$ on MMLU, trading accuracy for output diversity. Greedy MI instead stays near the top of the accuracy ranking while avoiding strongly correlated models: by $k = 5$, its ensembles span four model families, with all members above 72% accuracy on MEDMCQA and 79% on MMLU. The terms of (11) explain this. For the models Greedy MI adds on MEDMCQA, relevance minus redundancy is negative, so mRMR would score each addition as a loss of information; the error-dependence term outweighs this deficit and makes the net information gain positive. Model family is only a rough proxy for this dependence: on MEDMCQA, Mistral-3.1 and Mistral-3.2 are nearly redundant ($\rho = 0.89$), while GPT-4.1-mini and GPT-4.1-nano are almost uncorrelated ($\rho = 0.13$), which is why Greedy MI relies on measured dependence rather than family labels. Full results across temperatures, runs, and aggregation rules are in Appendices H and I.



(a) MEDMCQA (b) MMLU (c) AG News

(d) MEDMCQA (e) MMLU (f) AG News

Figure 3: Gaussian-copula validation across datasets (temp $= 0.7$, run 1). **Top row** (a–c): empirical vs. copula-modeled pairwise joint error probabilities $P(E_i \cap E_j)$; tight clustering around the diagonal confirms second-order fit. **Bottom row** (d–f): distribution of simultaneous errors; copula (orange) matches the observed tail (blue), confirming higher-order fit. The copula model captures error dependence consistently across binary (MEDMCQA, MMLU) and multi-class (AG News) settings.

5.2 Experiments on AG News Classification

We next test whether these patterns extend beyond binary tasks. AG News (Zhang et al., 2016) is a 4-class news classification dataset where we use all 7600 articles in its test set, each labeled World, Sports, Business, or Science/Technology. Unlike MEDMCQA and MMLU, which we convert into binary queries, we keep AG News in its original 4-class form. With more than two classes, two models can both be wrong yet predict different labels, so prediction redundancy and error dependence, which are closely tied in binary tasks, become distinct. AG News therefore gives a sharper test of the error-dependence term in (11).

Model pool. The AG News pool comprises 11 LLMs: Qwen3-235b and Qwen3-next-80b Yang et al. (2025a); Mistral-small-3.1-24b and Mistral-small-3.2-24b Mistral AI (2026); Gemini-2.5-flash and Gemini-2.5-flash-lite Comanici et al. (2025); GPT-4.1-mini, GPT-4.1-nano, and GPT-4o-mini OpenAI (2026); LLaMA-3.1-8b Grattafiori et al. (2024); and phi-4 Abdin et al. (2024). Individual accuracies span 73–88% ($\bar{\alpha} = 81.7\%$), with $\bar{\rho} = 0.91$, substantially higher than the binary settings. On AG News, forward selection uses a categorical naive Bayes classifier, which handles the 4-class model outputs directly, instead of logistic regression. The evaluation protocol also differs slightly: we use one run per temperature and ten random 90/10 splits per temperature, which again gives 30 evaluations. All other baselines are the same as in the binary experiments.

Copula validation. Figures 3c and f show the same diagnostics for AG News. Pairwise scatter clusters tightly along the diagonal across all $\binom{11}{2} = 55$ pairs, and the simultaneous-error histogram closely matches the empirical tail. The copula model captures error dependence in the multi-class setting as well as in the binary settings.

Results. Figure 2c shows test error versus ensemble size on AG News. Greedy MI achieves the lowest or tied-lowest mean test error at every k from 2 to 9, including the five diversity-based baselines, and its lead is clearest at $k = 4$–6. For example, it reaches 11.5% at $k = 5$ versus 11.8% for the strongest

Table 1: Cross-dataset comparison at ensemble budget $k = 5$. All selectors use MAP aggregation. Greedy MI achieves the lowest mean test error on all three datasets. Gains are reported in percentage points (pp). Oracle error denotes the minimum test error over all subsets of size k, averaged over the same evaluations.

Dataset	M	$\bar{\rho}$	k	**Best Baseline Error**	**Greedy MI Error**	**Gain (pp)**	**Oracle Error**
MEDMCQA	2	0.55	5	16.9% (LogReg/Kuncheva)	**16.3%**	0.6	16.0%
MMLU	2	0.60	5	14.8% (LogReg)	**14.1%**	0.7	13.8%
AG News	4	0.91	5	11.8% (LLM-TOPLA)	**11.5%**	0.3	11.0%

Table 2: Consistency check of Theorem 3.3 (stated for equal accuracies, equicorrelation, $K \to \infty$, and Naive Bayes aggregation), with the floor evaluated at the pool averages $\bar{\alpha}$ and $\bar{\rho}$, which are estimated from the training split. The empirical full-ensemble error ($k = K$, MAP aggregation) is close to this floor on MEDMCQA and MMLU. Greedy MI's mid-budget error lies below the full-pool floor because each selected subset has its own, lower operative floor.

Dataset	M	$\bar{\alpha}$	$\bar{\rho}$	**Floor (Thm. 3.3)**	**Empirical** ($k=K$)	**Greedy MI (mid-k)**
MEDMCQA	2	74.6%	0.55	18.6%	18.8%	**16.3%** ($k = 5$)
MMLU	2	77.3%	0.60	16.7%	17.1%	**14.1%** ($k = 5$)
AG News	4	81.7%	0.91	12.3%	15.1%	**11.5%** ($k = 5$)

baselines (LLM-TOPLA, Disagreement, and Kuncheva). The diversity-based selectors match Greedy MI at several budgets but do not remove its advantage in this moderate-budget range. The smaller gains than on the binary datasets are consistent with the high error correlation ($\bar{\rho} = 0.91$) and limited empirical improvement: no method improves on the best single model (11.8%) by more than 0.7 percentage points at any budget. At large k, error rises across methods, consistent with the MAP aggregator having to estimate 4^k prediction-pattern probabilities from finite data, and the curves converge as the selected subsets increasingly overlap (Appendix K).

Selection behavior. The pattern seen on MEDMCQA and MMLU repeats on AG News. Both mRMR and Greedy MI start with Qwen3-235b, the most accurate model. mRMR then adds LLaMA-3.1-8b (ranked 10th of 11 by accuracy) at $k = 2$ and phi-4 (ranked 11th) by $k = 5$, trading accuracy for output diversity. Greedy MI instead adds Gemini-2.5-flash, Gemini-2.5-flash-lite, GPT-4.1-mini, and GPT-4o-mini at $k = 2$–5, all above 79% accuracy, and skips the two Mistral models, the second and third most accurate, which an accuracy ranking would add next. Its picks include two Gemini and two OpenAI models, so Greedy MI does not simply spread across families; it follows the measured error dependence. Full results across temperatures and aggregation rules are in Appendix K.

Table 1 summarizes the results at $k = 5$: Greedy MI achieves the lowest mean test error among non-oracle selectors on all three datasets, spanning binary balanced (MEDMCQA, MMLU) and multi-class (AG News) settings. Its error is within 0.3 percentage points of the oracle on MEDMCQA and MMLU, and 0.5 percentage points on AG News. Gains over the strongest baselines are smaller in the high-correlation regime, consistent with Theorem 3.3.

Saturation validation. Table 2 is a consistency check rather than a test of Theorem 3.3, which assumes equal accuracies, equicorrelation, $K \to \infty$, and Naive Bayes aggregation: it compares the floor at the pool averages $(\bar{\alpha}, \bar{\rho})$ with the full-ensemble error ($k = K$) under MAP aggregation. On MEDMCQA and MMLU ($M = 2$, where MSC is vacuous), the two agree within 0.4 percentage points. On AG News, the empirical error (15.1%) exceeds the floor (12.3%): beyond the smaller pool and the 4^k pattern space, models that both err pick the same wrong class 94% of the time versus 33% under MSC, so the MSC floor underestimates the true floor. Greedy MI's mid-budget error lies below the full-pool floor because each selected subset has its own $(\bar{\alpha}, \bar{\rho})$: selection lowers the operative floor rather than breaking it.

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 correla-

tions ($\bar{\rho}=0.90$). Due to this very high accuracy and high correlation, improvements from ensemble selection are small and all methods perform comparably; full results are provided in Appendix J.

6 Conclusion

In this work, we studied budgeted LLM ensemble selection through an information-theoretic lens. Modeling LLM responses as M-ary symmetric channels coupled to a Gaussian copula on the latent error events captures both individual accuracies and inter-model correlations. Under independent errors, the optimal ensemble is the Top-k by accuracy; under correlated errors, accuracy alone is insufficient, and for Naive Bayes (plurality) aggregation of equally accurate, equicorrelated models a non-vanishing error floor emerges whose closed form depends only on the accuracy α, the correlation ρ, and M. Whether the optimal aggregator saturates under realistic errors remains open. Motivated by these results, we proposed *Greedy MI*, a greedy mutual information selection algorithm that favors accurate models whose errors are less dependent, which lowers the operative floor of the selected ensemble. On MEDMCQA, MMLU, and AG News, Greedy MI attains the lowest or tied-lowest mean test error among twelve selection methods at budgets of three to seven models, with gains of up to 0.7 percentage points over the strongest baseline, while on IMDB all methods perform comparably. On MEDMCQA and MMLU, the error of the full pool is consistent with the predicted floor.

Limitations and Discussion

Our framework targets classification with a finite label set, not open-ended generation. While the Gaussian-copula model captures key correlation patterns regarding when errors occur, our theoretical M-ary extension relies on the M-ary Symmetric Channel (MSC) assumption, which posits that incorrect predictions are distributed uniformly and independently among the wrong classes. In practice, LLMs often exhibit skewed, semantically driven failure modes where models reinforce the exact same specific incorrect prediction. MSC is therefore unrealistic as a model of LLM errors, and we use it only in our theoretical analysis. While our empirical Greedy MI algorithm robustly bypasses this limitation by computing the full conditional mutual information directly from data (thereby capturing these asymmetric semantic failures), relaxing the theoretical MSC assumption to formally model correlated specific failure modes presents a promising direction for future work. Finally, the M^k growth of the pattern space limits the algorithm's scalability to very large budgets ($k \geq 10$), where increased estimation variance allows simpler, independence-assuming baselines to match the performance of information-theoretic selection.

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.

Acknowledgments and Disclosure of Funding

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

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.

Banfield, R. E., Hall, L. O., Bowyer, K. W., and Kegelmeyer, W. Ensemble diversity measures and their application to thinning. *Information Fusion*, 6(1):49–62, 2005. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2004.04.005. URL https://www.sciencedirect.com/science/article/pii/S1566253504000387. Diversity in Multiple Classifier Systems.

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. An information theoretic perspective on multiple classifier systems. In *Multiple Classifier Systems (MCS)*, volume 5519 of *Lecture Notes in Computer Science*, pp. 344–353. Springer, 2009.

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.

Caruana, R., Niculescu-Mizil, A., Crew, G., and Ksikes, A. Ensemble selection from libraries of models. In *Proceedings of the Twenty-First International Conference on Machine Learning*, ICML '04, pp. 18, New York, NY, USA, 2004. Association for Computing Machinery. ISBN 1581138385. doi: 10.1145/1015330.1015432. URL https://doi.org/10.1145/1015330.1015432.

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.

Cunningham, P. and Carney, J. Diversity versus quality in classification ensembles based on feature selection. In López de Mántaras, R. and Plaza, E. (eds.), *Machine Learning: ECML 2000*, pp. 109–116, Berlin, Heidelberg, 2000. Springer Berlin Heidelberg. ISBN 978-3-540-45164-8.

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=0ju2Qu9jvn.

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.

Gu, J., Jiang, X., Shi, Z., Tan, H., Zhai, X., Xu, C., Li, W., Shen, Y., Ma, S., Liu, H., Wang, S., Zhang, K., Wang, Y., Gao, W., Ni, L., and Guo, J. A survey on llm-as-a-judge, 2025. URL https://arxiv.org/abs/2411.15594.

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/.

Inan, H., Upasani, K., Chi, J., Rungta, R., Iyer, K., Mao, Y., Tontchev, M., Hu, Q., Fuller, B., Testuggine, D., and Khabsa, M. Llama guard: Llm-based input-output safeguard for human-ai conversations, 2023. URL https://arxiv.org/abs/2312.06674.

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/.

Kuncheva, L. I. and Whitaker, C. J. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Mach. Learn.*, 51(2):181–207, May 2003. ISSN 0885-6125. doi: 10.1023/A:1022859003006. URL https://doi.org/10.1023/A:1022859003006.

Laban, P., Schnabel, T., Bennett, P. N., and Hearst, M. A. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177, 2022. doi: 10.1162/tacl_a_00453. URL https://aclanthology.org/2022.tacl-1.10/.

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.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., and Lowe, R. Training language models to follow instructions with human feedback, 2022. URL https://arxiv.org/abs/2203.02155.

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.

Proakis, J. and Salehi, M. *Digital Communications*. McGraw-Hill International Edition. McGraw-Hill, 2008. ISBN 9780071263788. URL https://books.google.com.tr/books?id=ksh0GgAACAAJ.

Purucker, L., Schneider, L., Anastacio, M., Beel, J., Bischl, B., and Hoos, H. Q(d)o-es: Population-based quality (diversity) optimisation for post hoc ensemble selection in automl, 2023. URL https://arxiv.org/abs/2307.08364.

Shayegh, B., Lee, H. H.-B., Zhu, X., Cheung, J. C. K., and Mou, L. Error diversity matters: An error-resistant ensemble method for unsupervised dependency parsing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(23):25119–25127, April 2025. ISSN 2159-5399. doi: 10.1609/aaai.v39i23.34697. URL http://dx.doi.org/10.1609/aaai.v39i23.34697.

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/.

Thorne, J., Vlachos, A., Christodoulopoulos, C., and Mittal, A. FEVER: a large-scale dataset for fact extraction and VERification. In Walker, M., Ji, H., and Stent, A. (eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 809–819, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1074. URL https://aclanthology.org/N18-1074/.

Turkmen, Y., Buyukates, B., and Bastopcu, M. Balancing information accuracy and response timeliness in networked LLMs. In *IEEE INFOCOM 2026 - IEEE Conference on Computer Communications*, pp. 1–9, 2026.

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.

Wood, D., Mu, T., Webb, A. M., Reeve, H. W., Lujan, M., and Brown, G. A unified theory of diversity in ensemble learning. *Journal of machine learning research*, 24(359):1–49, 2023.

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.

Yates, R. D. and Goodman, D. J. *Probability and stochastic processes: a friendly introduction for electrical and computer engineers*. John Wiley & Sons, 2014.

Zhang, W., Deng, Y., Liu, B., Pan, S., and Bing, L. Sentiment analysis in the era of large language models: A reality check. In Duh, K., Gomez, H., and Bethard, S. (eds.), *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 3881–3906, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-naacl.246. URL https://aclanthology.org/2024.findings-naacl.246/.

Zhang, X., Zhao, J., and LeCun, Y. Character-level convolutional networks for text classification, 2016. URL https://arxiv.org/abs/1509.01626.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., Zhang, H., Gonzalez, J. E., and Stoica, I. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL https://openreview.net/forum?id=uccHPGDlao.

Appendix

Table 3: Organization of the Appendix.

Section	Contents
A	Related Work
B	Proof of Theorem 3.1 (independence optimality)
C	Proof of Theorem 3.2 (MI decomposition)
D	Proof of Theorem 3.3 (saturation limit)
E	Saturation Floor under Imbalanced Priors
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	Additional Experiments on AG News
L	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. (2026) 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.

Ensemble Selection and Diversity: Classical ensemble learning studies both model selection and the relationship between diversity and predictive performance. Caruana et al. (2004) construct ensembles through greedy selection guided by validation performance. Cunningham & Carney (2000) introduce an entropy-based diversity measure, while Kuncheva & Whitaker (2003) evaluate pairwise and non-pairwise measures and question their reliability as predictors of ensemble accuracy. More recent work explores population-based quality-diversity optimization for post hoc selection Purucker et al. (2023) and error-diversity-aware selection for unsupervised dependency parsing Shayegh et al. (2025). Wood et al. (2023) provide a theoretical account of diversity through loss-dependent bias–variance–diversity relationships. Our work studies cross-LLM error dependence using a Gaussian copula and connects this dependence to mutual information and budgeted subset selection.

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 $Y \in \mathcal{M}$ and output $X \in \mathcal{M}$ specified by the transition kernel:

$$\mathbb{P}(X = y' \mid Y = y) = \begin{cases} 1-\epsilon, & y' = y, \\ \dfrac{\epsilon}{M-1}, & y' \neq y. \end{cases} \tag{14}$$

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:*

$$\delta = \frac{\epsilon_2 - \epsilon_1}{1 - \dfrac{M\,\epsilon_1}{M-1}}. \tag{15}$$

Proof. Let $Y \in \mathcal{M}$ 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:

$$\begin{aligned} \mathbb{P}(Z_2 = y \mid Y = y) &= \mathbb{P}(Z_2 = y \mid Z_1 = y)\mathbb{P}(Z_1 = y \mid Y = y) \\ &\quad + \sum_{y'' \neq y} \mathbb{P}(Z_2 = y \mid Z_1 = y'')\mathbb{P}(Z_1 = y'' \mid Y = y) \\ &= (1-\delta)(1-\epsilon_1) + (M-1) \cdot \frac{\delta}{M-1} \cdot \frac{\epsilon_1}{M-1} \\ &= (1-\delta)(1-\epsilon_1) + \frac{\delta\,\epsilon_1}{M-1}. \end{aligned} \tag{16}$$

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

$$\begin{aligned} 1 - \epsilon_1 - \delta + \delta\,\epsilon_1 + \frac{\delta\,\epsilon_1}{M-1} &= 1 - \epsilon_2 \\ -\epsilon_1 - \delta\left[1 - \epsilon_1\left(1 + \tfrac{1}{M-1}\right)\right] &= -\epsilon_2 \\ \delta\left[1 - \frac{M\,\epsilon_1}{M-1}\right] &= \epsilon_2 - \epsilon_1. \end{aligned} \tag{17}$$

This directly yields (15). 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 $\epsilon_1 \leq \epsilon_2 \leq \cdots \leq \epsilon_K$, 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 K independent MSCs where channel j has error rate $\epsilon_j < (M-1)/M$, indexed such that $\epsilon_1 \leq \epsilon_2 \leq \cdots \leq \epsilon_K$, and

an arbitrary class prior $\boldsymbol{\pi}$ with full support on $\mathcal{M}$. For any subset $\mathcal{S} \subseteq \mathcal{K}$ with $|\mathcal{S}| = k$, the top-k set H_k satisfies:

$$I(Y; \mathbf{X}_{\mathbf{H_k}}) \geq I(Y; \mathbf{X}_{\mathcal{S}}), \tag{18}$$

$$\mathbb{P}_e(H_k) \leq \mathbb{P}_e(\mathcal{S}). \tag{19}$$

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

We now define a bijection $\pi : \mathcal{S} \to H_k$ satisfying $\epsilon_{\pi(j)} \leq \epsilon_j$. To see this, first set $\pi(j) = j$ for all $j \in \mathcal{S} \cap H_k$. The remaining elements satisfy $|\mathcal{S} \setminus H_k| = |H_k \setminus \mathcal{S}|$, so we can bijectively map $\mathcal{S} \setminus H_k$ onto $H_k \setminus \mathcal{S}$. For any $j \in \mathcal{S} \setminus H_k$ 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 $\widetilde{X}_{\mathcal{S}} = (\widetilde{X}_j)_{j \in \mathcal{S}}$ as follows: for each $j \in \mathcal{S}$, 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 $\widetilde{\mathbf{X}}_{\mathcal{S}} \overset{d}{=} \mathbf{X}_{\mathcal{S}}$: 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 $\widetilde{X}_{\mathcal{S}}$ matches that of $\mathbf{X}_{\mathcal{S}}$, i.e.,

$$\widetilde{\mathbf{X}}_{\mathcal{S}} \overset{d}{=} \mathbf{X}_{\mathcal{S}}. \tag{20}$$

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

$$Y \longrightarrow \mathbf{X}_{\mathbf{H_k}} \longrightarrow \widetilde{\mathbf{X}}_{\mathcal{S}} \overset{d}{=} \mathbf{X}_{\mathcal{S}}. \tag{21}$$

Applying the Data Processing Inequality Cover & Thomas (2006) to the Markov chain in Equation (21), we obtain:

$$I(Y; \mathbf{X}_{\mathcal{S}}) = I(Y; \widetilde{\mathbf{X}}_{\mathcal{S}}) \leq I(Y; \mathbf{X}_{\mathbf{H_k}}), \tag{22}$$

which proves (18).

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

$$\mathbb{P}_e(\mathcal{S}) = \min_{\hat{Y}} \mathbb{P}(\hat{Y}(\mathbf{X}_{\mathcal{S}}) \neq Y) = \min_{\hat{Y}} \mathbb{P}(\hat{Y}(\widetilde{\mathbf{X}}_{\mathcal{S}}) \neq Y). \tag{23}$$

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

$$\mathbb{P}_e(H_k) = \min_{\hat{Y}} \mathbb{P}(\hat{Y}(\mathbf{X}_{\mathbf{H_k}}) \neq Y) \leq \mathbb{P}_e(\mathcal{S}), \tag{24}$$

which proves (19). 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). Both inequalities in Theorem 3.1 hold with equality when $\mathcal{S}$ consists of k channels with the smallest error rates (ties allowed). For any other $\mathcal{S}$, at least one component is strictly degraded ($\delta_j > 0$ for some j), so the mutual-information inequality (6) is strict whenever all error rates are positive. The error-probability inequality (7) need not be strict, because the MAP rule may ignore a weaker channel. For example, with $M = 2$, a balanced prior, and $\boldsymbol{\epsilon} = (0.1, 0.2, 0.4)$, the MAP rule on $\{1, 2\}$ and on $\{1, 3\}$ follows channel 1 whenever the two disagree, so both subsets have error probability exactly 0.1, while their mutual information differs (0.636 vs. 0.542 bits). Similarly, with prior $(0.9, 0.1)$ and $k = 1$, the MAP rule ignores a single channel with $\epsilon = 0.3$ or $\epsilon = 0.4$ and always outputs the more likely class, giving error probability 0.1 in both cases.

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 amounts of information about Y. For example, under the prior $\boldsymbol{\pi} = (0.9, 0.05, 0.05)$ and error rate $\epsilon = 0.2$ for every class, the symmetric channel carries 0.190 bits about Y, whereas a channel that sends every error on class 1 to class 2 (and is symmetric otherwise) carries 0.275 bits. Concentrated errors are more predictable and hence more informative; under a uniform prior, the symmetric channel is in fact the least informative among all channels with the same per-class accuracy, since averaging any such channel over relabelings of the classes yields the symmetric channel and mutual information is convex in the channel. 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 $K = 3$ 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 $\mathcal{S} = \{1, 3\}$.
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 (15):

- $\delta_1 = 0$: 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:

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

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

C Proof of Theorem 3.2

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:

$$I(X;Y \mid Z) \triangleq H(X \mid Z) - H(X \mid Y,Z) = \mathbb{E}_Z\left[D_{\mathrm{KL}}(P_{X,Y|Z} \| P_{X|Z}P_{Y|Z})\right]. \tag{25}$$

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:*

$$I(X;Y,Z) = I(X;Z) + I(X;Y \mid Z) = I(X;Y) + I(X;Z \mid Y). \tag{26}$$

Lemma C.3 (Entropy of M-ary Errors)**.** *Let $Y \in \mathcal{M}$ be the true label, $X \in \mathcal{M}$ 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:*

$$H(X \mid Y) = H(E \mid Y) + \epsilon \log(M-1). \tag{27}$$

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 3.2) and then provide its proof.

Theorem 3.2 [Accuracy-Redundancy-Error Decomposition] For outputs $X_j \in \mathcal{M}$ and true label $Y \in \mathcal{M}$ 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 $\mathcal{S}$ as:

$$\Delta(j \mid \mathcal{S}) \triangleq I(Y; X_j \mid \mathbf{X}_{\mathcal{S}}). \tag{28}$$

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

$$\Delta(j \mid \mathcal{S}) = \underbrace{I(Y;X_j)}_{\text{Accuracy}} - \underbrace{I(X_j;\mathbf{X}_{\mathcal{S}})}_{\text{Redundancy}} + \underbrace{I(E_j;E_{\mathcal{S}})}_{\text{Error Correlation}} + \underbrace{\Lambda_j(\mathcal{S})}_{\text{Correction}}, \tag{29}$$

where the label-dependent correction term is given by:

$$\Lambda_j(\mathcal{S}) \triangleq I(E_j; Y \mid E_{\mathcal{S}}) - I(E_j;Y). \tag{30}$$

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:

$$\Delta(j \mid \mathcal{S}) = I(Y; X_j \mid \mathbf{X}_{\mathcal{S}}) = H(X_j \mid \mathbf{X}_{\mathcal{S}}) - H(X_j \mid \mathbf{X}_{\mathcal{S}}, Y). \tag{31}$$

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

$$H(X_j \mid \mathbf{X}_{\mathcal{S}}) = H(X_j) - I(X_j;\mathbf{X}_{\mathcal{S}}). \tag{32}$$

Next, we evaluate the term $H(X_j \mid \mathbf{X}_{\mathcal{S}}, Y)$. In our Gaussian-copula framework, the correlation among models is fully captured by the latent error indicators $E_{\mathcal{S}}$. 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 $\mathbf{X}_\mathcal{S}$ and Y perfectly determines $E_\mathcal{S}$, applying the logic from Lemma C.3 to the conditioned variables yields:

$$H(X_j \mid \mathbf{X}_\mathcal{S}, Y) = H(E_j \mid E_\mathcal{S}, Y) + \epsilon_j \log(M-1). \tag{33}$$

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

$$H(E_j \mid E_\mathcal{S}, Y) = H(E_j \mid Y) - I(E_j; E_\mathcal{S} \mid Y). \tag{34}$$

Substituting (32), (33), and (34) into (31), we have:

$$\Delta(j \mid \mathcal{S}) = H(X_j) - I(X_j; \mathbf{X}_\mathcal{S}) - H(E_j \mid Y) + I(E_j; E_\mathcal{S} \mid Y) - \epsilon_j \log(M-1). \tag{35}$$

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:

$$I(Y; X_j) = H(X_j) - H(X_j \mid Y) = H(X_j) - \left[H(E_j \mid Y) + \epsilon_j \log(M-1)\right]. \tag{36}$$

Rearranging to solve for $H(X_j)$:

$$H(X_j) = I(Y; X_j) + H(E_j \mid Y) + \epsilon_j \log(M-1). \tag{37}$$

Substituting Equation (37) into Equation (35), observe that the error dispersion terms $\epsilon_j \log(M-1)$ perfectly cancel out:

$$\begin{aligned}\Delta(j \mid \mathcal{S}) = &\Big[I(Y; X_j) + H(E_j \mid Y) + \epsilon_j \log(M-1)\Big] \\ &- I(X_j; \mathbf{X}_\mathcal{S}) - H(E_j \mid Y) + I(E_j; E_\mathcal{S} \mid Y) - \epsilon_j \log(M-1) \\ = &\, I(Y; X_j) - I(X_j; \mathbf{X}_\mathcal{S}) + I(E_j; E_\mathcal{S} \mid Y).\end{aligned} \tag{38}$$

Applying Lemma C.2 to $I(E_j; E_\mathcal{S}, Y)$ in two ways:

$$I(E_j; E_\mathcal{S}, Y) = I(E_j; E_\mathcal{S}) + I(E_j; Y \mid E_\mathcal{S}), \tag{39}$$

$$I(E_j; E_\mathcal{S}, Y) = I(E_j; Y) + I(E_j; E_\mathcal{S} \mid Y). \tag{40}$$

Equating Equation (39) and Equation (40) and solving for $I(E_j; E_\mathcal{S} \mid Y)$:

$$I(E_j; E_\mathcal{S} \mid Y) = I(E_j; E_\mathcal{S}) + \underbrace{\left[I(E_j; Y \mid E_\mathcal{S}) - I(E_j; Y)\right]}_{\Lambda_j(\mathcal{S})}. \tag{41}$$

Substituting Equation (41) into the canceled result from Equation (38):

$$\Delta(j \mid \mathcal{S}) = I(Y; X_j) - I(X_j; \mathbf{X}_\mathcal{S}) + I(E_j; E_\mathcal{S}) + \Lambda_j(\mathcal{S}). \tag{42}$$

This completes the proof. □

Remark C.4 (Interpretation of Terms). The decomposition in Equation (29) has an intuitive interpretation:

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

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

$$I(E_j; Y) = 0 \quad \textit{and} \quad I(E_j; Y \mid E_\mathcal{S}) = 0, \quad \forall j, S. \tag{43}$$

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

$$\Delta(j \mid \mathcal{S}) = I(Y; X_j) - I(X_j; \mathbf{X}_\mathcal{S}) + I(E_j; E_\mathcal{S}). \tag{44}$$

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

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

$$-I(E_j; Y) \leq \Lambda_j(\mathcal{S}) \leq H(Y) - I(E_j; Y). \tag{45}$$

Proof. By non-negativity of mutual information, $I(E_j; Y \mid E_{\mathcal{S}}) \geq 0$. Therefore:

$$\Lambda_j(\mathcal{S}) = I(E_j; Y \mid E_{\mathcal{S}}) - I(E_j; Y) \geq -I(E_j; Y). \tag{46}$$

Conditional mutual information is bounded by conditional entropy:

$$I(E_j; Y \mid E_{\mathcal{S}}) \leq H(Y \mid E_{\mathcal{S}}) \leq H(Y). \tag{47}$$

Therefore:

$$\Lambda_j(\mathcal{S}) = I(E_j; Y \mid E_{\mathcal{S}}) - I(E_j; Y) \leq H(Y) - I(E_j; Y). \tag{48}$$

□

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.,

$$I(E_j, E_{\mathcal{S}}; Y) \leq \delta, \tag{49}$$

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

D Proof of Theorem 3.3

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

$$T_j \triangleq \sqrt{\rho}\, U + \sqrt{1-\rho}\, V_j, \qquad j = 1, 2, \ldots \tag{50}$$

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

Define the error indicator by thresholding:

$$E_j \triangleq \mathbb{1}(T_j \leq \tau). \tag{51}$$

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 $\mathcal{M} \setminus \{y\}$, with the wrong-label draws conditionally independent across j given $(E_1, \ldots, E_K)$ 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

$$\mathbb{P}(E_j = 1) = \mathbb{P}(T_j \leq \tau) = \Phi(\tau) = 1 - \alpha. \tag{52}$$

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

D.2 Proof of Theorem 3.3

Conditioned on $U = u$, we have from (50) 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

$$p(u) \triangleq \mathbb{P}(E_j = 1 \mid U = u) = \mathbb{P}(T_j \leq \tau \mid U = u) = \Phi\left(\frac{\tau - \sqrt{\rho}\, u}{\sqrt{1-\rho}}\right). \tag{53}$$

For each class $y \in \mathcal{M}$, 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 NumbersYates & Goodman (2014) applied conditional on (U, Y), we have almost surely

$$\bar{N}_K(y^\star) \xrightarrow[K\to\infty]{a.s.} 1 - p(U), \qquad \bar{N}_K(y) \xrightarrow[K\to\infty]{a.s.} \frac{p(U)}{M-1} \text{ for each } y \neq y^\star. \tag{54}$$

Under the balanced prior, the decision rule reduces to plurality voting: $\widehat{Y}_{\mathrm{NB}} = \arg\max_{y \in \mathcal{M}} \bar{N}_K(y)$, with ties broken arbitrarily. From (54), 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

$$1 - p(U) > \frac{p(U)}{M-1} \quad \Longleftrightarrow \quad p(U) < \frac{M-1}{M}. \tag{55}$$

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

$$\mathbb{1}\{\widehat{Y}_{\mathrm{NB}} \neq Y\} \;\xrightarrow[K\to\infty]{a.s.}\; \mathbb{1}\left\{p(U) > \frac{M-1}{M}\right\}. \tag{56}$$

Taking expectations yields

$$\lim_{K\to\infty} \mathbb{P}(\widehat{Y}_{\mathrm{NB}} \neq Y) = \mathbb{P}\left(p(U) > \frac{M-1}{M}\right). \tag{57}$$

Now let us find the probability of the event $p(U) > (M-1)/M$ explicitly. From (53), 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

$$\mathbb{P}\left(p(U) > \frac{M-1}{M}\right) = \Phi\left(\frac{\tau - \sqrt{1-\rho}\, \Phi^{-1}((M-1)/M)}{\sqrt{\rho}}\right). \tag{58}$$

Substituting $\tau = \Phi^{-1}(1-\alpha)$ using (52) in (58) gives

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

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 $\lim_{K\to\infty} \mathbb{P}(\widehat{Y}_{\mathrm{NB}} \neq Y)$ vanishes, recovering the independent-error regime. As ρ increases to 1, the argument approaches $\Phi^{-1}(1-\alpha)$, which implies $\lim_{K\to\infty} \mathbb{P}(\widehat{Y}_{\mathrm{NB}} \neq Y) = 1-\alpha$, 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 Saturation Floor under Imbalanced Priors

Corollary E.1 (Prior-Invariance of the NB Saturation Floor)**.** *Under the same Gaussian-copula model as in Theorem 3.3, but with an arbitrary class prior $\boldsymbol{\pi}$ on $\mathcal{M}$ satisfying $\pi_y > 0$ for all y, the asymptotic probability of error satisfies*

$$\lim_{K\to\infty} \mathbb{P}(\widehat{Y}_{\mathrm{NB}} \neq Y) = \Phi\left(\frac{\Phi^{-1}(1-\alpha) - \sqrt{1-\rho}\,\Phi^{-1}((M-1)/M)}{\sqrt{\rho}}\right), \tag{59}$$

which is identical to the balanced-prior expression in Theorem 3.3.

The proof of Corollary E.1 is provided below. Intuitively, the NB 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. We now extend the proof of Theorem 3.3 to an arbitrary class prior $\boldsymbol{\pi}$ on $\mathcal{M}$ with full support, i.e., $\pi_y > 0$ for all $y \in \mathcal{M}$.

E.1 Proof of Corollary E.1

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

$$\widehat{Y}_{\mathrm{NB}} = \arg\max_{y\in\mathcal{M}} \{\log \pi_y + \ell_K(y)\}, \tag{60}$$

where $\ell_K(y) \triangleq \sum_{j=1}^K \log \mathbb{P}(X_j \mid Y = y)$ treats the outputs as conditionally independent given Y, which is what defines the NB rule (under the copula model, they are conditionally independent only given both Y and U). Conditional on $(U, Y) = (u, y^\star)$, the empirical fractions $\bar{N}_K(y)$ converge as in (54), and the MSC kernel (1) gives

$$\ell_K(y) = \sum_{j=1}^K \mathbb{1}(X_j = y)\log(1-\epsilon) + \sum_{j=1}^K \mathbb{1}(X_j \neq y)\log\frac{\epsilon}{M-1}, \tag{61}$$

where $\epsilon = 1 - \alpha$. Substituting the empirical fractions and using (54), the conditional log-likelihood per model converges almost surely:

$$\frac{1}{K}\ell_K(y) \xrightarrow[K\to\infty]{a.s.} L(y; U, y^\star), \tag{62}$$

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

$$L(y; U, y^\star) = \begin{cases} (1-p(U))\,\log(1-\epsilon) + p(U)\,\log\dfrac{\epsilon}{M-1}, & y = y^\star, \\ \dfrac{p(U)}{M-1}\,\log(1-\epsilon) + \left(1 - \dfrac{p(U)}{M-1}\right)\log\dfrac{\epsilon}{M-1}, & y \neq y^\star. \end{cases} \tag{63}$$

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 (63), the asymptotic per-sample log-likelihood gap between the truth and any alternative is

$$L(y^\star; U, y^\star) - L(y; U, y^\star) = \left(1 - p(U) - \frac{p(U)}{M-1}\right)\log\frac{(1-\epsilon)(M-1)}{\epsilon}, \qquad y \neq y^\star. \tag{64}$$

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 $p(U) \to (M-1)/M$.

Returning to the MAP rule (60), the difference between the score for $y^\star$ and any alternative y is

$$\big[\log \pi_{y^\star} + \ell_K(y^\star)\big] - \big[\log \pi_y + \ell_K(y)\big] = \underbrace{\log \frac{\pi_{y^\star}}{\pi_y}}_{\text{finite offset}} + \underbrace{\ell_K(y^\star) - \ell_K(y)}_{=K\cdot[L(y^\star;U,y^\star)-L(y;U,y^\star)]+o(K)}. \tag{65}$$

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 (54) place mass below $1/M$ on $y^\star$ and above $1/M$ on every other class, so the gap in (64) is negative and the rule fails to recover $y^\star$.

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

$$\mathbb{1}\{\widehat{Y}_{\mathrm{NB}} \neq Y\} \xrightarrow[K\to\infty]{a.s.} \mathbb{1}\left\{p(U) > \frac{M-1}{M}\right\}, \tag{66}$$

which is identical to (56) in the balanced case. Taking expectations and applying the same calculation as in (58) yields

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

which is the same expression as in Theorem 3.3. This completes the proof.

Remark E.2 (Rate of convergence). While the asymptotic floor in Corollary E.1 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 (65) 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 NB 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 3.3, 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 $Y \in \mathcal{M}$ be the true label drawn from an arbitrary class prior $\boldsymbol{\pi}$, and let $\mathbf{X}_{\mathcal{K}} = \{X_1, \ldots, X_K\}$ 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 $\mathcal{S} \subseteq \mathcal{K}$, model outputs are conditionally independent given (Y, D):*

$$\mathbb{P}(\mathbf{X}_{\mathcal{S}} \mid Y, D) = \prod_{j \in \mathcal{S}} \mathbb{P}(X_j \mid Y, D). \tag{67}$$

Then, for every $\mathcal{S} \subseteq \mathcal{K}$,

$$I(Y; \mathbf{X}_{\mathcal{S}}) = I(Y; \mathbf{X}_{\mathcal{S}} \mid D) - I(Y; D \mid \mathbf{X}_{\mathcal{S}}). \tag{68}$$

We refer to $I(Y; \mathbf{X}_{\mathcal{S}} \mid D)$ 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 $I(Y; D \mid \mathbf{X}_{\mathcal{S}})$ as the Price of Difficulty which measures the remaining ambiguity that is related to not observing D after seeing the ensemble.

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

$$I(Y; D \mid \mathbf{X}_{\mathcal{S}}) = H(Y \mid \mathbf{X}_{\mathcal{S}}) - H(Y \mid \mathbf{X}_{\mathcal{S}}, D), \tag{69}$$

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 $\mathcal{M}$ 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 $I(Y; \mathbf{X}_{\mathcal{S}} \mid D)$ that is monotone submodular, hence greedy achieves a $(1 - 1/e)$ approximation guarantee for maximizing this term under $|\mathcal{S}| \leq k$ 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 $I(Y; D \mid \mathbf{X}_{\mathcal{S}})$. If this term does not vanish as $|\mathcal{S}|$ grows, then adding more models cannot fully remove uncertainty about Y, which provides an information-theoretic explanation for the saturation phenomenon characterized in Theorem 3.3. 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 $\mathcal{M}$ 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 $F(\mathcal{S}) \triangleq I(Y; \mathbf{X}_{\mathcal{S}} \mid D)$.

Let's first derive the decomposition in Equation (68). We can apply the chain rule for mutual information to the triple $(Y, \mathbf{X}_{\mathcal{S}}, D)$ in two different ways:

$$I(Y; \mathbf{X}_{\mathcal{S}}, D) = I(Y; \mathbf{X}_{\mathcal{S}}) + I(Y; D \mid \mathbf{X}_{\mathcal{S}}), \tag{70}$$

$$I(Y; \mathbf{X}_{\mathcal{S}}, D) = I(Y; D) + I(Y; \mathbf{X}_{\mathcal{S}} \mid D). \tag{71}$$

Equating Equations (70) and (71) yields

$$I(Y; \mathbf{X}_{\mathcal{S}}) = I(Y; \mathbf{X}_{\mathcal{S}} \mid D) + I(Y; D) - I(Y; D \mid \mathbf{X}_{\mathcal{S}}). \tag{72}$$

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 (72) gives $I(Y;\mathbf{X}_{\mathcal{S}})=I(Y;\mathbf{X}_{\mathcal{S}} \mid D) - I(Y;D \mid \mathbf{X}_{\mathcal{S}})$, which is exactly Equation (68).

Now, we can derive Equation (69) by using the definition of conditional mutual information,

$$I(Y;D \mid \mathbf{X}_{\mathcal{S}})=H(Y \mid \mathbf{X}_{\mathcal{S}}) - H(Y \mid \mathbf{X}_{\mathcal{S}}, D), \tag{73}$$

which matches Equation (69).

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

$$F(\mathcal{S})=I(Y;\mathbf{X}_{\mathcal{S}} \mid D)=H(\mathbf{X}_{\mathcal{S}} \mid D) - H(\mathbf{X}_{\mathcal{S}} \mid Y, D). \tag{74}$$

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

$$H(\mathbf{X}_{\mathcal{S}} \mid Y, D)=\sum_{j\in\mathcal{S}} H(X_j \mid Y, D), \tag{75}$$

so the set function $\mathcal{S} \mapsto H(\mathbf{X}_{\mathcal{S}} \mid Y, D)$ 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 $\mathcal{S}\mapsto H(\mathbf{X}_{\mathcal{S}} \mid D)$ is submodular (see, e.g., Cover & Thomas (2006)). These submodularity and modularity properties hold for any discrete alphabet, in particular $\mathcal{M}$ for any $M\geq 2$. Recalling the identity $F(\mathcal{S})$ in (74), we conclude that F is submodular as a submodular function minus a modular one. Finally, for any $j\notin\mathcal{S}$,

$$F(S\cup\{j\}) - F(\mathcal{S})=I(Y;X_j \mid \mathbf{X}_{\mathcal{S}}, D)\geq 0, \tag{76}$$

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 and diversity-based selection 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 for the binary case. 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 3.3). 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 Consistent 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 joint distribution directly with Laplace smoothing
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

▷ Derive marginal distributions strictly via summation
for $a \in \mathcal{A}$ **do**
  $\hat{P}(A=a) \leftarrow \sum_{b\in\mathcal{B}} \hat{P}(A=a, B=b)$
end for
for $b \in \mathcal{B}$ **do**
  $\hat{P}(B=b) \leftarrow \sum_{a\in\mathcal{A}} \hat{P}(A=a, B=b)$
end for

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

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

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 $\mathcal{S}_{\text{Shapley}}$ of size k

Initialize $v_j \leftarrow 0$ for all $j \in \{1, \ldots, K\}$
$U_{\text{full}} \leftarrow \text{MAP_Error}(Y^{\text{tr}}, X^{\text{tr}}, Y^{\text{tr}}, X^{\text{tr}})$
for $p = 1$ to P **do**
  $\pi \leftarrow \text{RandomPermutation}(K)$
  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
    $\mathcal{S}_{\text{curr}} \leftarrow \pi[1 \ldots \text{pos}]$
    curr_error $\leftarrow \text{MAP_Error}(Y^{\text{tr}}, X^{\text{tr}}_{S_{\text{curr}}}, Y^{\text{tr}}, X^{\text{tr}}_{S_{\text{curr}}})$
    $v_j \leftarrow v_j + (\text{prev_error} - \text{curr_error})$
    prev_error $\leftarrow$ curr_error
  end for
end for
$v_j \leftarrow v_j / P$ for all $j \in \{1, \ldots, K\}$
$\mathcal{S}_{\text{Shapley}} \leftarrow$ 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.

Naive Bayes Forward Selection: This baseline utilizes a greedy forward-selection algorithm structurally identical to the Logistic Regression approach, but replaces the evaluation model with a

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 $\mathcal{S}_{\text{LogReg}}$
$\mathcal{S}_{\text{LogReg}} \leftarrow \emptyset$
Convert labels: $Y_{\text{lr}}^{\text{tr}} \leftarrow (Y^{\text{tr}} + 1)/2$
for $\mathcal{S}tep = 1$ to k **do**
  best_acc $\leftarrow -\infty$, best_model $\leftarrow -1$
  for $j \in \{1, \ldots, K\} \setminus S_{\text{LogReg}}$ **do**
    $\mathcal{S}_{\text{cand}} \leftarrow S_{\text{LogReg}} \cup \{j\}$
    $\text{acc}_{\text{cand}} \leftarrow$ 5-Fold-CV-Accuracy$(X_{S_{\text{cand}}}^{\text{tr}}, Y_{\text{lr}}^{\text{tr}}, \text{model} = \text{LogisticRegression})$
    if $\text{acc}_{\text{cand}} >$ best_acc **then**
      best_acc $\leftarrow \text{acc}_{\text{cand}}$
      best_model $\leftarrow j$
    end if
  end for
  $\mathcal{S}_{\text{LogReg}} \leftarrow S_{\text{LogReg}} \cup \{\text{best_model}\}$
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 $\mathcal{S}_{\text{TOPLA}}$
best_score $\leftarrow -\infty$, $\mathcal{S}_{\text{TOPLA}} \leftarrow \emptyset$
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: $F \leftarrow \mathbb{1}(X_c^{\text{tr}} \neq Y^{\text{tr}})$
  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
  $U_{\text{acc}} \leftarrow 1 -$ MAP_Error$(Y^{\text{tr}}, X_c^{\text{tr}}, Y^{\text{tr}}, X_c^{\text{tr}})$
  score $\leftarrow 0.6 \cdot \lambda + 0.4 \cdot U_{\text{acc}}$
  if score $>$ best_score **then**
    best_score $\leftarrow$ score
    $\mathcal{S}_{\text{TOPLA}} \leftarrow c$
  end if
end for

Categorical Naive Bayes classifier (Algorithm 6). At each iterative step up to budget k, the algorithm temporarily adds each unselected candidate to the current ensemble.

Algorithm 6 Forward Selection via Naive Bayes (with 5-Fold CV)

Require: Training outputs $\mathbf{X}_{\mathcal{K}}$, target labels Y, target subset size k
Ensure: Selected subset $\mathcal{S}$
Initialize subset $\mathcal{S} \leftarrow \emptyset$
for $t = 1$ **to** k **do**
  Initialize $best_score \leftarrow -\infty$
  Initialize $best_candidate \leftarrow \emptyset$
  for each candidate model $j \in \mathcal{K} \setminus \mathcal{S}$ **do**
    Form candidate subset $\mathcal{S}_{temp} \leftarrow S \cup \{j\}$
    $cv_score \leftarrow$ Evaluate $\mathbf{X}_{S_{temp}}$ using 5-Fold CV Naive Bayes
    if $cv_score > best_score$ **then**
      $best_score \leftarrow cv_score$
      $best_candidate \leftarrow j$
    end if
  end for
  $\mathcal{S} \leftarrow S \cup \{best_candidate\}$
end for
return $\mathcal{S}$

G.3 Diversity-Based Selection Algorithms

We include five diversity-based baselines: Society Entropy Cunningham & Carney (2000), Disagreement Kuncheva & Whitaker (2003), Kohavi–Wolpert (KW) Variance Kuncheva & Whitaker (2003), Kuncheva diversity Kuncheva & Whitaker (2003), and the Percentage Correct Diversity Measure (PCDM) Banfield et al. (2005). They use the same greedy selection procedure, with a different diversity score for each baseline. All quantities below are computed from the training split $(X^{\mathrm{tr}}, Y^{\mathrm{tr}})$.

Let $\mathcal{K} = \{1, \ldots, K\}$ be the model pool and let k be the desired subset size. The empirical accuracy of model j is

$$\hat{p}_j = \frac{1}{N_{\mathrm{tr}}} \sum_{i=1}^{N_{\mathrm{tr}}} \mathbb{1}\left(X_{ij}^{\mathrm{tr}} = Y_i^{\mathrm{tr}}\right).$$

For a candidate subset $\mathcal{S}$ of size $r = |\mathcal{S}|$, define the numbers of correct and incorrect predictions on example i as

$$C_i(\mathcal{S}) = \sum_{j \in \mathcal{S}} \mathbb{1}\left(X_{ij}^{\mathrm{tr}} = Y_i^{\mathrm{tr}}\right), \qquad W_i(\mathcal{S}) = r - C_i(\mathcal{S}).$$

The following scores apply to both binary and multi-class tasks. We set every diversity score to zero when $r \leq 1$.

Society Entropy: This score measures how spread out the predicted labels are within the subset. It is high when the models assign their votes to several classes, regardless of which label is correct. Let $\mathcal{Y}$ be the label set and define

$$q_{iy}(\mathcal{S}) = \frac{1}{r} \sum_{j \in \mathcal{S}} \mathbb{1}\left(X_{ij}^{\mathrm{tr}} = y\right), \qquad y \in \mathcal{Y}.$$

For $r \geq 2$, the implemented score is

$$D_{\mathrm{SE}}(\mathcal{S}) = -\frac{1}{N_{\mathrm{tr}}} \sum_{i=1}^{N_{\mathrm{tr}}} \sum_{y \in \mathcal{Y}} q_{iy}(\mathcal{S}) \log q_{iy}(\mathcal{S}),$$

where $0 \log 0 = 0$. We use the natural logarithm, as in the code.

Disagreement: Our implementation measures disagreement between correctness indicators: one model is correct while another is incorrect. Its normalization corresponds to drawing two members independently with replacement from the subset:

$$D_{\mathrm{Dis}}(\mathcal{S}) = \frac{1}{N_{\mathrm{tr}} r^2} \sum_{i=1}^{N_{\mathrm{tr}}} 2 C_i(\mathcal{S}) W_i(\mathcal{S}), \qquad r \geq 2.$$

For multi-class tasks, two different incorrect labels still have the same correctness indicator.

Kohavi–Wolpert Variance: This score measures the variance of the subset's binary correctness indicators on each example, averaged over the training split. It is largest when similar numbers of models are correct and incorrect:

$$D_{\mathrm{KW}}(\mathcal{S}) = \frac{1}{N_{\mathrm{tr}} r^2} \sum_{i=1}^{N_{\mathrm{tr}}} C_i(\mathcal{S}) W_i(\mathcal{S}), \qquad r \geq 2.$$

Hence $D_{\mathrm{Dis}} = 2 D_{\mathrm{KW}}$: Disagreement with $\eta = 0.5$ selects exactly the same subsets as KW variance with $\eta = 1.0$, so the two baselines differ only in their other grid point. This is expected, since the averaged disagreement and the KW variance are proportional (Kuncheva & Whitaker, 2003).

Kuncheva Diversity: This score counts the smaller of the correct and incorrect groups on each example. We normalize by the largest possible size of that smaller group, so the score lies between zero and one:

$$D_{\mathrm{Kun}}(\mathcal{S}) = \frac{1}{N_{\mathrm{tr}}\left(r - \lceil r/2 \rceil\right)} \sum_{i=1}^{N_{\mathrm{tr}}} \min\{C_i(\mathcal{S}), W_i(\mathcal{S})\}, \qquad r \geq 2.$$

PCDM: PCDM counts the fraction of training examples on which the proportion of correct models is between 0.1 and 0.9, including both endpoints. It rewards subsets whose members are neither almost all correct nor almost all incorrect:

$$D_{\mathrm{PCDM}}(\mathcal{S}) = \frac{1}{N_{\mathrm{tr}}} \sum_{i=1}^{N_{\mathrm{tr}}} \mathbb{1}\left(0.1 \leq \frac{C_i(\mathcal{S})}{r} \leq 0.9\right), \qquad r \geq 2.$$

Shared Selection Procedure: For each diversity score D, we greedily maximize

$$F_\eta(\mathcal{S}) = \sum_{j \in \mathcal{S}} \hat{p}_j + \eta D(\mathcal{S}), \qquad \eta \in \{0.5, 1.0\}.$$

The first term is the sum of individual model accuracies, rather than the accuracy of the aggregated ensemble. The scores have different ranges ($D_{\mathrm{KW}} \leq 1/4$, $D_{\mathrm{Dis}} \leq 1/2$, $D_{\mathrm{SE}} \leq \log M$, and $D_{\mathrm{Kun}}, D_{\mathrm{PCDM}} \leq 1$), so the same η gives each score a different effective weight. We use η for the diversity weight to distinguish it from the smoothing parameter α used in MI estimation.

For each budget k, we compare the two resulting subsets using training error under the baseline's aggregation rule $g \in \{\mathrm{MAP}, \mathrm{MV}, \mathrm{WV}\}$. The subset with the smaller error is retained. For MAP, this error is measured on the same training split used to fit the aggregator; the weight is not selected by cross-validation. Test labels are used only for the final evaluation. Candidate ties are resolved by the smallest model index, and ties between the two weights retain $\eta = 0.5$.

Algorithm 7 Greedy Selection with an Accuracy–Diversity Score

Require: Training data $(X^{\mathrm{tr}}, Y^{\mathrm{tr}})$, model pool $\mathcal{K}$, budget k, diversity score D, aggregation rule g
Ensure: Selected subset $\mathcal{S}_{D,g}$ of size k

```
Compute training accuracies (p̂_j)_{j∈K}
e_best ← +∞, S_{D,g} ← ∅
for η in the ordered list (0.5, 1.0) do
    S_η ← ∅
    for t = 1 to k do
        j* ← arg max_{j∈K\S_η} F_η(S_η ∪ {j}) {Smallest index on ties}
        S_η ← S_η ∪ {j*}
    end for
    e_η ← Training_Error_g(Y^tr, X^tr_{S_η})
    if e_η < e_best then
        e_best ← e_η, S_{D,g} ← S_η
    end if
end for
return S_{D,g}
```

In the implementation, each greedy path is computed once per metric and weight, and its prefixes are reused across budgets. Repeated diversity scores and aggregation errors are cached. These steps reduce repeated computation without changing the selection criterion.

G.4 Aggregation Methods

Algorithm 8 summarizes our MAP aggregator. For a selected subset $\mathcal{S}$ of size k, we estimate $P(Y \mid X_{\mathcal{S}})$ from the training split by counting labels for each prediction pattern $x_{\mathcal{S}} \in \{-1, +1\}^k$ (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 $X_{\mathcal{S}}$. Although our empirical implementations are illustrated primarily for binary tasks, both the Greedy MI selection and the likelihood-based aggregation algorithms naturally extend to the M-ary multi-class setting without structural changes, requiring only an expansion of the categorical state-space in the empirical joint distributions.

Algorithm 8 The MAP Aggregation

Require: Training labels $Y^{\text{tr}} \in \{-1, +1\}^{N_{\text{tr}}}$, training predictions $X_{\mathcal{S}}^{\text{tr}} \in \{-1, +1\}^{N_{\text{tr}} \times k}$
Require: Test predictions $X_{\mathcal{S}}^{\text{te}} \in \{-1, +1\}^{N_{\text{te}} \times k}$
Ensure: Predicted labels $\hat{Y}^{\text{te}} \in \{-1, +1\}^{N_{\text{te}}}$
Map each $\mathbf{x}_{\mathcal{S}} \in \{-1, +1\}^k$ to index $\text{idx}(\mathbf{x}_{\mathcal{S}}) \in \{1, \ldots, 2^k\}$: ▷ Index prediction patterns
  $\text{idx}(\mathbf{x}_{\mathcal{S}}) \leftarrow 1 + \sum_{j=1}^{k} \frac{x_j+1}{2} \cdot 2^{j-1}$

▷ 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 $k = |\mathcal{S}|$. 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 9 implements the standard majority voting aggregation. Given predictions from the selected subset $\mathcal{S}$, we sum the binary votes $X_j \in \{-1, +1\}$ and predict the sign of the total. When the ensemble size $|\mathcal{S}|$ is odd, ties cannot occur; for even $|\mathcal{S}|$, 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 9 Majority Voting (MV) Aggregation

Require: Selected subset $\mathcal{S}$, model predictions $\mathbf{X}_{\mathcal{S}} = (X_j)_{j \in \mathcal{S}}$
$V \leftarrow \sum_{j \in \mathcal{S}} X_j$ ▷ 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 10 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.

Algorithm 10 Weighted Majority Voting (WV) Aggregation

Require: Selected subset $\mathcal{S}$, model predictions $\mathbf{X}_\mathcal{S} = (X_j)_{j\in\mathcal{S}}$, accuracies $(p_j)_{j\in\mathcal{S}}$

for $j \in \mathcal{S}$ **do**
  $w_j \leftarrow \log \frac{p_j}{1-p_j}$ {Log-odds weight} ▷ Compute weight
end for
$V \leftarrow \sum_{j\in\mathcal{S}} w_j \cdot X_j$ ▷ Compute weighted vote
$\hat{Y}_{\text{WV}} \leftarrow \text{sign}(V)$

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 $\mathbb{P}(H_i = Y) = p_i$.*

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

Multi-Class Majority Voting: For standard Majority Voting (Algorithm 11), the ensemble evaluates a test instance by allowing each selected model to cast a single, equal-weight vote for its predicted class. The class that accumulates the highest total number of votes (the statistical mode) is selected as the final ensemble prediction, with ties broken arbitrarily.

Algorithm 11 Multi-Class Majority Voting (MV) Aggregation

Require: Selected subset $\mathcal{S}$, test predictions $X_{\mathcal{S}}^{\text{te}} \in \{1, \ldots, M\}^{N_{\text{te}} \times |\mathcal{S}|}$, class set $\mathcal{C} = \{1, \ldots, M\}$
Ensure: Predicted labels $\hat{Y}_{\text{MV}}^{\text{te}} \in \{1, \ldots, M\}^{N_{\text{te}}}$
for $i = 1$ to N_{te} **do**
  for $c \in \mathcal{C}$ **do**
    $V_c \leftarrow 0$ ▷ Initialize vote counts for this instance
  end for
  ▷ Tally the votes from the ensemble
  for $j \in \mathcal{S}$ **do**
    $c^* \leftarrow X_{i,j}^{\text{te}}$ ▷ Prediction of model j for instance i
    $V_{c^*} \leftarrow V_{c^*} + 1$ ▷ Cast one equal vote for class c^*
  end for
  ▷ Select the class with the most votes
  $\hat{Y}_{\text{MV},i}^{\text{te}} \leftarrow \arg\max_{c \in \mathcal{C}} V_c$ {Break ties arbitrarily}
end for

Multi-Class Weighted Majority Voting: Unlike binary classification where a single global accuracy defines a model's reliability, multi-class environments require tracking specialized expertise. We adapt standard Weighted Majority Voting by calculating a class-specific weight matrix $W_{j,c}$ (Algorithm 12). For every model j and class c, we compute the empirical precision of the model when predicting that specific class on the training set (applying Laplace smoothing to prevent numerical instability). The final weight is the log-odds of this precision. During inference, if a model predicts class c, it contributes its specific log-odds weight $W_{j,c}$ to that class's total tally, and the ensemble outputs the class with the maximum accumulated score.

Algorithm 12 Multi-Class Weighted Majority Voting (WMV) Aggregation

Require: Training labels Y^{tr}, training predictions $X_{\mathcal{S}}^{\text{tr}}$, test predictions $X_{\mathcal{S}}^{\text{te}}$, class set $\mathcal{C}$, smoothing α
Ensure: Predicted label $\hat{Y}_{\text{WMV}}$
Phase 1: Compute Class-Specific Weights
for $j \in \mathcal{S}$ **do**
  for $c \in \mathcal{C}$ **do**
    $N_{\text{pred}} \leftarrow \sum \mathbb{I}(X_j^{\text{tr}} = c)$ ▷ Count times model j predicted class c
    $N_{\text{correct}} \leftarrow \sum \mathbb{I}(X_j^{\text{tr}} = c \wedge Y^{\text{tr}} = c)$ ▷ Count times this prediction was correct
    $\text{Prec}_{j,c} \leftarrow \frac{N_{\text{correct}} + \alpha}{N_{\text{pred}} + \alpha|\mathcal{C}|}$ ▷ Compute Laplace-smoothed precision
    $W_{j,c} \leftarrow \log \frac{\text{Prec}_{j,c}}{1 - \text{Prec}_{j,c}}$ ▷ Class-specific log-odds weight
  end for
end for
Phase 2: Aggregate Votes
for $c \in \mathcal{C}$ **do**
  $V_c \leftarrow 0$ ▷ Initialize accumulated score for class c
end for
for $j \in \mathcal{S}$ **do**
  $c^* \leftarrow X_j^{\text{te}}$ ▷ Class predicted by model j for the test instance
  $V_{c^*} \leftarrow V_{c^*} + W_{j,c^*}$ ▷ Add model's specific weight for its predicted class
end for
$\hat{Y}_{\text{WMV}} \leftarrow \arg\max_{c \in \mathcal{C}} V_c$ ▷ Output the class with the maximum weighted vote

G.5 Gaussian-Copula Fitting and Sampling Algorithm

Algorithm 13 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 13 Gaussian-Copula Fitting and Synthetic Generation

Require: Model predictions $X \in \{-1, 1\}^{N \times K}$, 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 $E \in \{0, 1\}^{N \times K}$ where $E_{ij} \leftarrow \mathbb{1}(X_{ij} \neq Y_i)$
for $j = 1$ to K **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 $\mathbf{\Sigma} \leftarrow \mathbf{I}_{K \times K}$
for $j = 1$ to K **do**
  for $k = j + 1$ to K **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 $Z \in \mathbb{R}^{N \times K} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$
Generate labels $Y_{\text{syn}} \sim \text{Uniform}(\{-1, 1\})^N$
Initialize $X_{\text{syn}} \in \mathbb{R}^{N \times K}$
for $i = 1$ to N **do**
  for $j = 1$ to K **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":

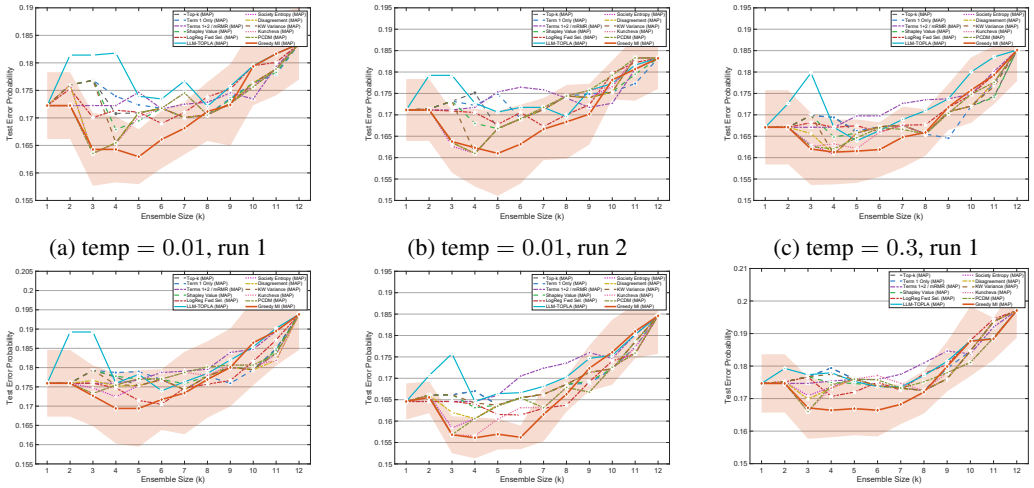
Example Query for MEDMCQA Dataset

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 4 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

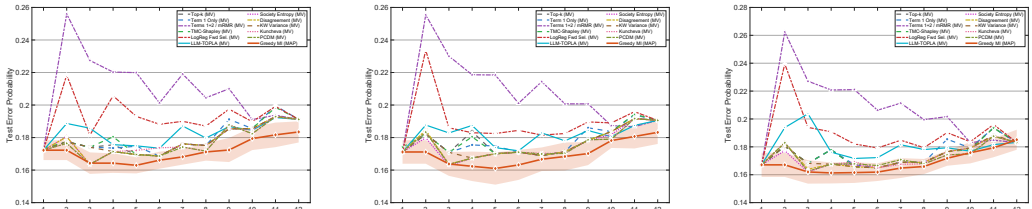
Figure 4: Experiments across temperatures and runs (**the MAP aggregation**) on MEDMCQA. Each panel corresponds to one temperature setting and random run. Shaded region represents the standard deviation.

H.2 Alternative Aggregation Rules

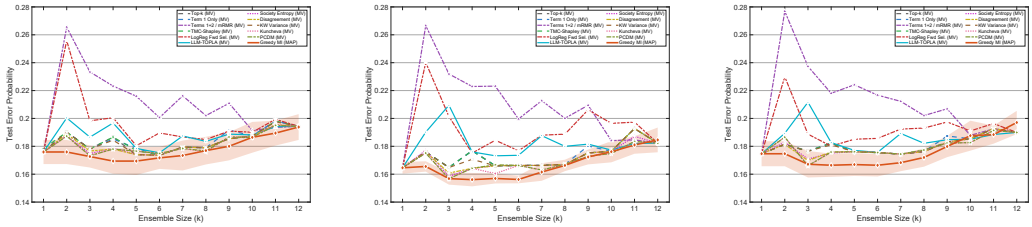
Our main experiments use the MAP aggregator for the selected subset $\mathcal{S}$ (Algorithm 8), since it directly estimates $\mathbb{P}(Y \mid X_{\mathcal{S}})$ and can exploit higher-order dependencies in correlated outputs. To

probe robustness to the aggregation rule, we additionally evaluate two widely-used alternatives for the baseline selectors: (i) *majority vote* (MV) (Algorithm 9), and (ii) *weighted majority vote* (WV) (Algorithm 10), 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.

From Figure 5 and Table 5, we observe that the mRMR selector (Terms 1+2) performs poorly under MV: at $k = 2$, its error reaches 0.264, versus 0.181 for Top-k and Term 1, and 0.171 for Greedy MI (MAP). This pattern 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 achieves the lowest or tied-lowest displayed mean error (e.g., $k = 5$: 0.163 vs. 0.170 for Top-k; $k = 6$: 0.164 vs. 0.170 for Top-k). Among the diversity-based baselines, Kuncheva and PCDM match Greedy MI at $k = 3$ (0.164), but the best diversity baseline at $k = 4$ yields 0.171, compared to 0.163 for Greedy MI.

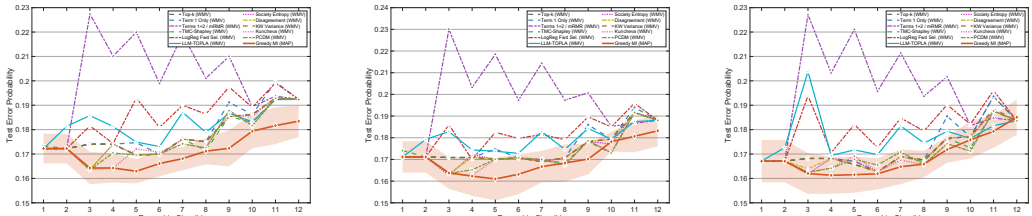


(a) temp $= 0.01$, run 1 (b) temp $= 0.01$, run 2 (c) temp $= 0.3$, run 1

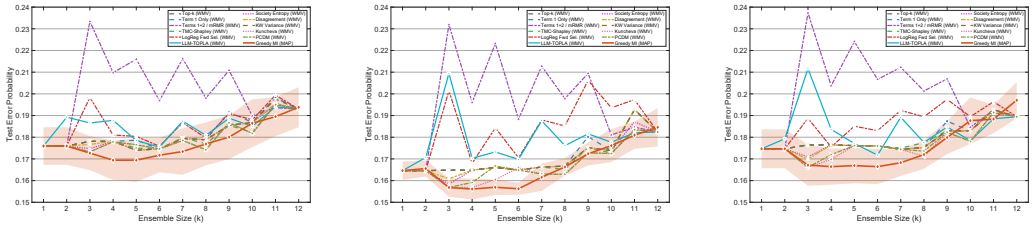


(d) temp $= 0.3$, run 2 (e) temp $= 0.7$, run 1 (f) temp $= 0.7$, run 2

Figure 5: MEDMCQA 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

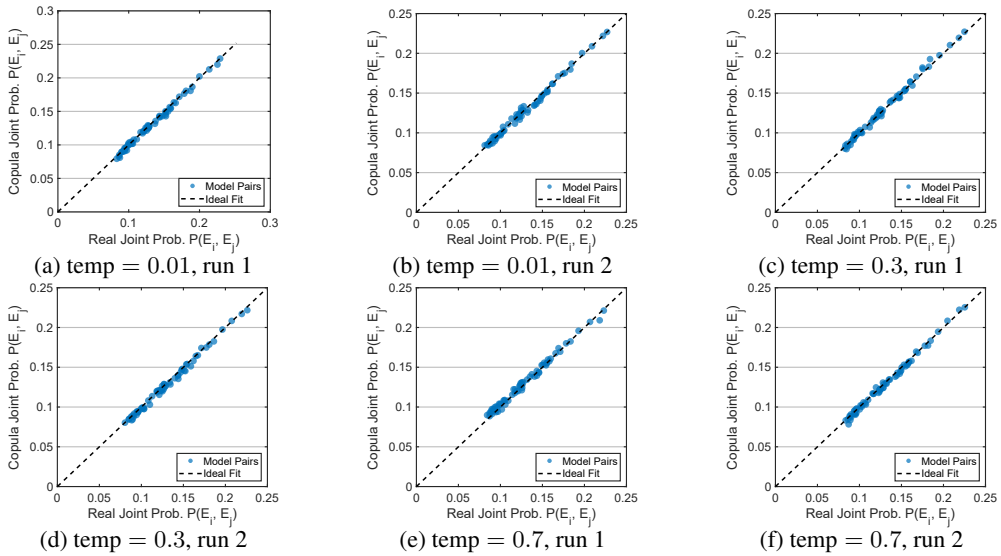
Figure 6: MEDMCQA results using **Weighted Ensemble (WE)** aggregation across temperatures and runs. Shaded region represents the standard deviation.

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 7 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.

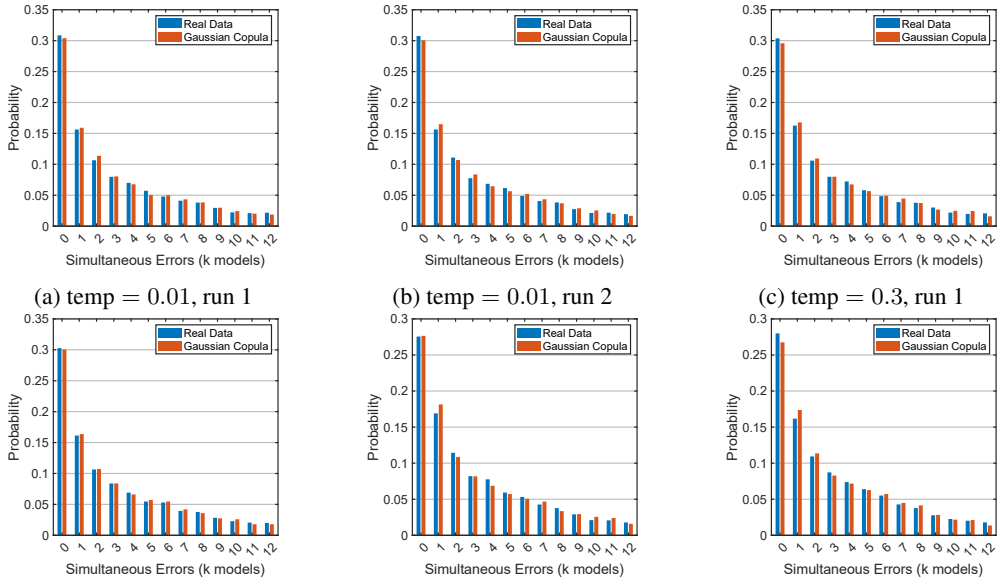
Figure 8 evaluates higher-order structure by showing the distribution of simultaneous errors that is, how many models fail on the same instance. The copula-generated distributions (orange bars) match the empirical frequencies (blue bars) well across all settings.



(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 7: 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 8: 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":

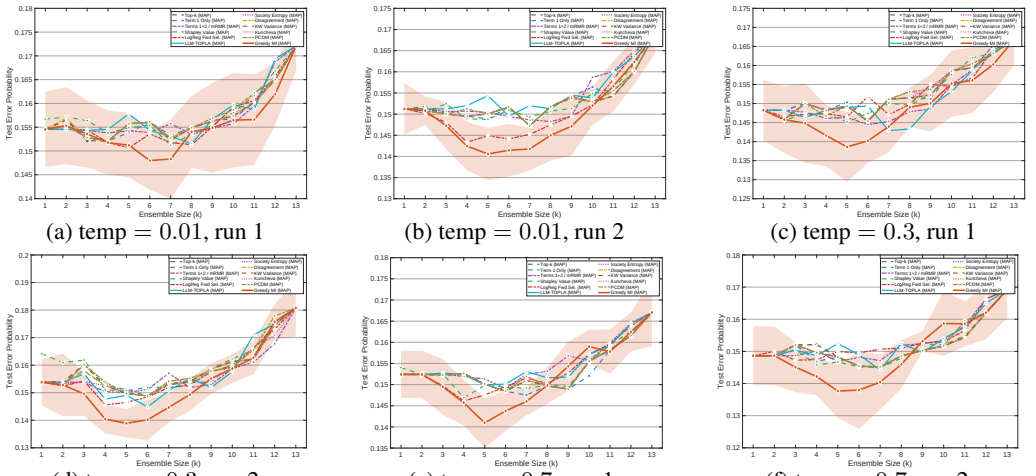
Example Query for MMLU Dataset

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 9 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

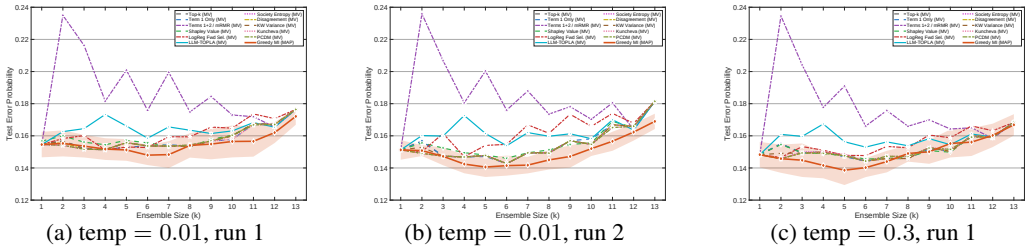
(d) temp $= 0.3$, run 2 (e) temp $= 0.7$, run 1 (f) temp $= 0.7$, run 2

Figure 9: 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.

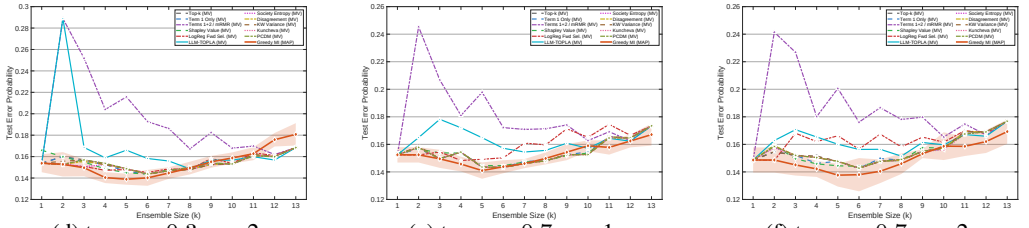
I.2 Experiments with Alternative Aggregation Rules

In this part, we follow the same approach as in Appendix H.2. From Figure 10 and Table 9, we observe that mRMR performs poorly under MV at small budgets: at $k = 2$, its error reaches 0.247. LLM-TOPLA also struggles, reaching 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. At moderate budgets, Greedy MI

maintains a lower mean error (e.g., $k = 4$: 0.144 versus 0.151 for Top-k and LogReg, and 0.168 for LLM-TOPLA; $k = 6$: 0.142 versus 0.145 for Top-k and 0.147 for TMC-Shapley). The five diversity-based baselines also remain above Greedy MI in this range, all yielding 0.151 at $k = 4$ and 0.145 at $k = 6$.

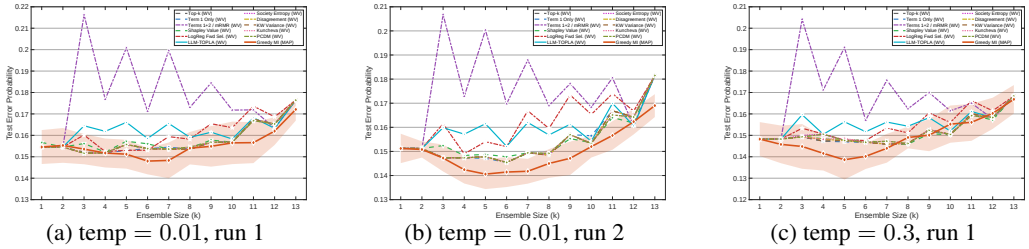


(a) temp $= 0.01$, run 1 (b) temp $= 0.01$, run 2 (c) temp $= 0.3$, run 1

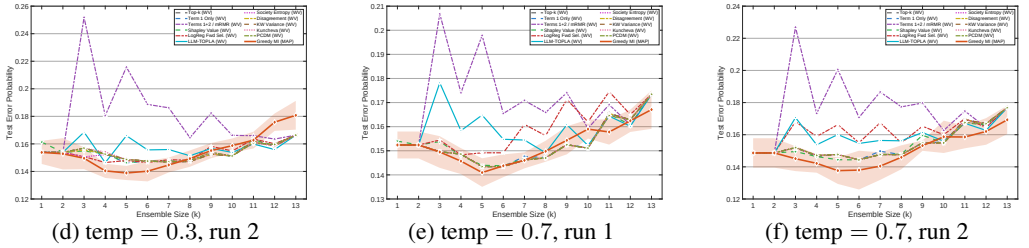


(d) temp $= 0.3$, run 2 (e) temp $= 0.7$, run 1 (f) temp $= 0.7$, run 2

Figure 10: MMLU results using **the 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 11: MMLU results using **Weighted Ensemble (WE)** aggregation across temperatures and runs. Shaded region represents the standard deviation.

Under WV, Figure 11 and Table 10 show that Terms 1+2 remains substantially worse in the mid-budget range (e.g., $k = 4$: 0.175; $k = 6$: 0.170), whereas Greedy MI (MAP) achieves 0.144 ± 0.007 at $k = 4$ and 0.142 ± 0.008 at $k = 6$. The diversity-based baselines also remain above Greedy MI at moderate budgets (e.g., all five yield 0.147 at $k = 6$), although they outperform it at $k = 10$ (0.153 versus 0.157).

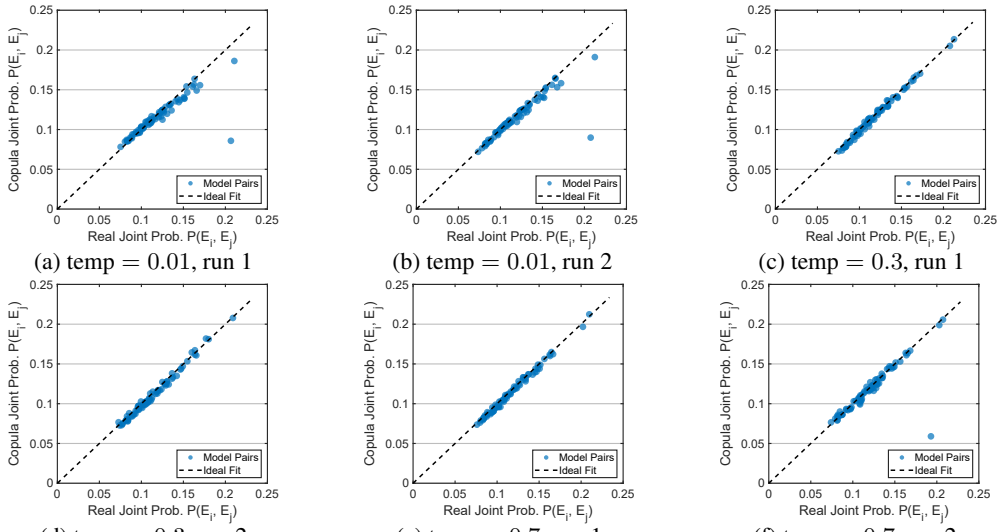
At larger budgets, WV can narrow or reverse the gap for simpler selectors: for example, Top-k reaches 0.153 at $k = 10$, compared to 0.157 for Greedy MI (MAP). However, this advantage is not consistent across budgets: at $k = 11$, Greedy MI achieves 0.158, compared to 0.165 for Top-k.

Overall, changing the aggregation rule can improve the relative performance of Top-k/Term 1 at some larger budgets, but it does not resolve the poor performance of mRMR under MV/WV. Greedy MI with MAP remains most competitive in the moderate-budget regime ($k \approx 3$–7). The per-setting curves also show instances where baselines with WV aggregation surpass Greedy MI with MAP. Such reversals may reflect limitations of greedy selection and the increasing difficulty of estimating high-dimensional joint distributions as k grows.

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 12 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.

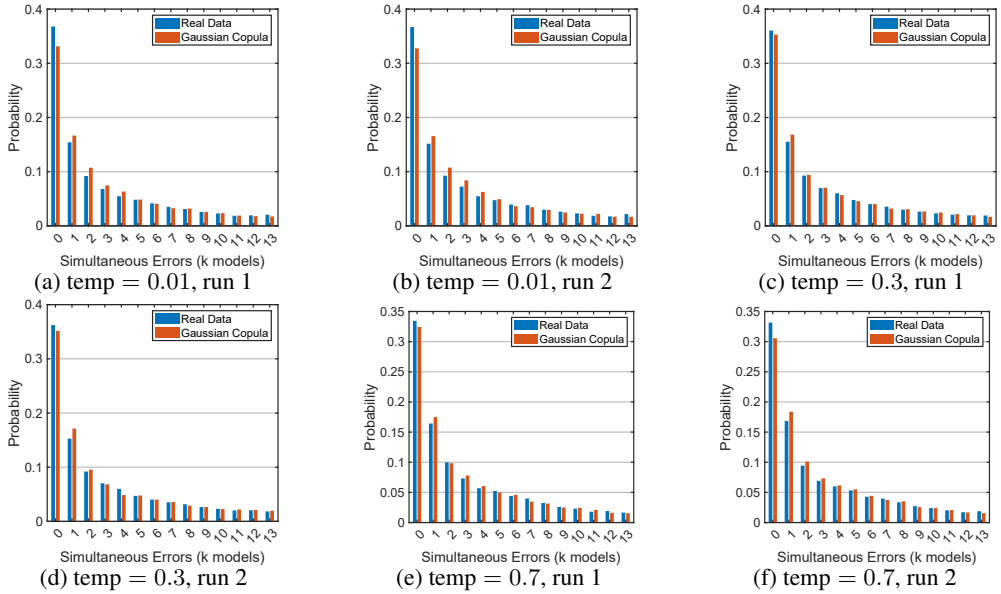
Figure 13 examines higher-order dependencies by plotting the distribution of simultaneous errors that is, the number of models that makes an error on the same instance. This diagnostic is critical for ensemble performance: if all models fail together on hard examples, no aggregation rule can recover. 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 12: 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 13: 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 26 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.

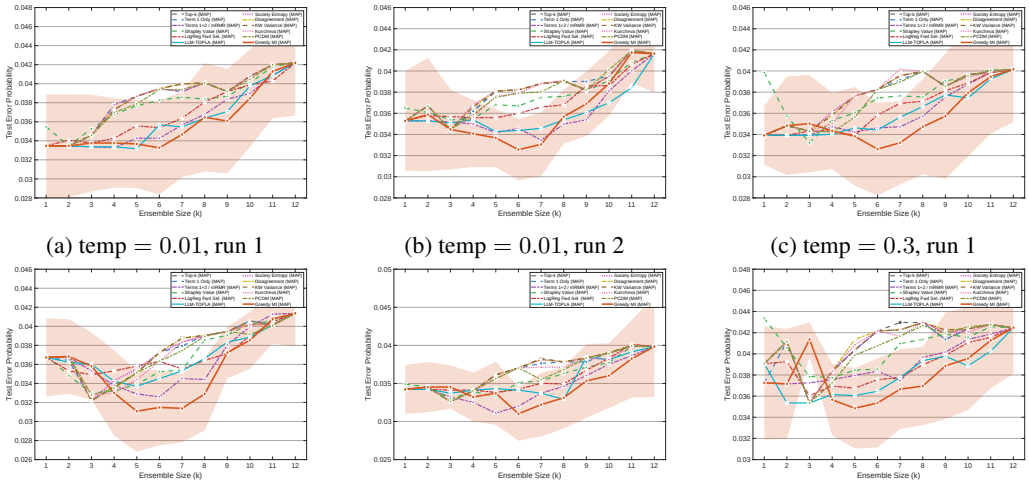
User: Review: The scriptwriters, directors, and actors have lost sight of the cornerstone of a good story - the concept of suspension of disbelief. In Volcano, the concept goes up in smoke almost as quickly as the city. Contrary to earlier commentators, I much preferred Dantes Peak amongst the 97 vintage 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 14 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 3.3 because when correlations are near-uniform and high, the error floor $\Phi(\Phi^{-1}(1-\alpha)/\sqrt{\bar{\rho}})$ (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

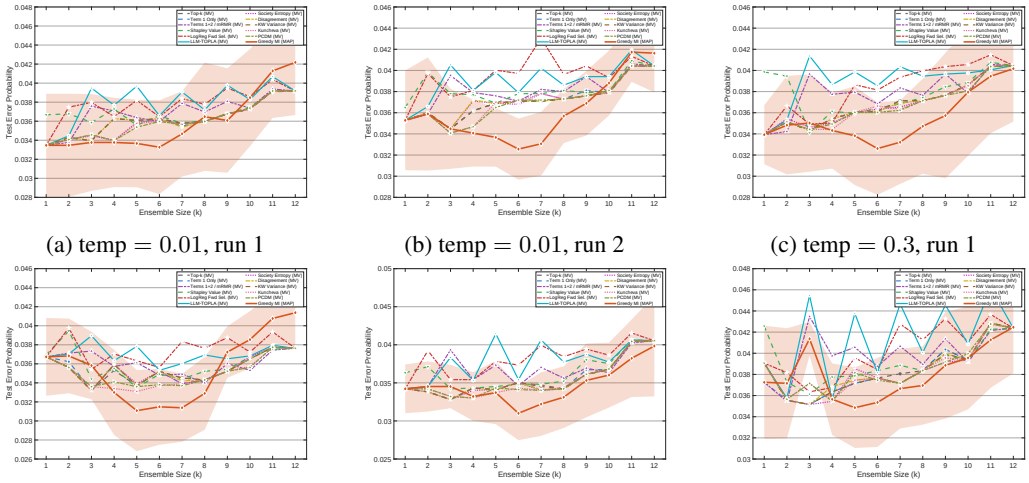
Figures 15 and 16 and Tables 13 and 14 show MV and WV ablations for IMDB. Unlike MEDMCQA and MMLU, all methods perform comparably here, i.e., the curves largely overlap and differences fall within the variance bands.



(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: 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.



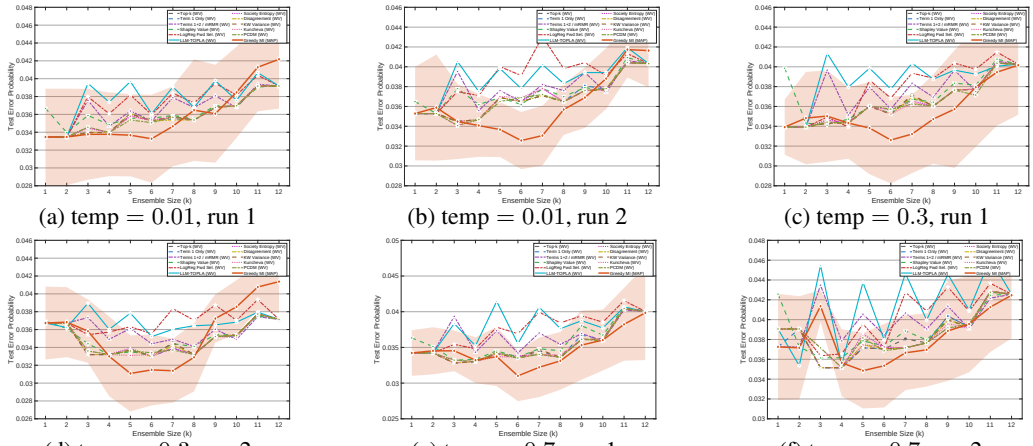
(a) temp $= 0.01$, run 1 (b) temp $= 0.01$, run 2 (c) temp $= 0.3$, run 1

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

Figure 15: IMDB movie reviews dataset results using **Majority Vote (MV)** aggregation across temperatures and runs. Shaded region represents the standard deviation.

J.3 Copula Validation

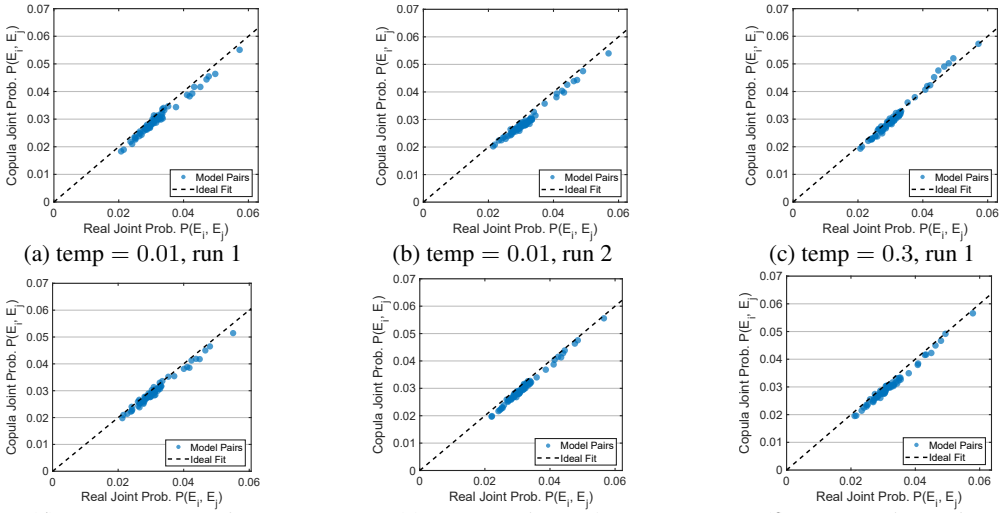
Figure 17 shows the pairwise fit remains tight along the diagonal across all settings. Figure 18 reveals that the error distribution is heavily concentrated at $k = 0$ (reflecting the high model accuracies) with the copula accurately capturing this structure. The near-absence of simultaneous errors ($k \geq 2$) further explains why ensemble selection yields limited gains on 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

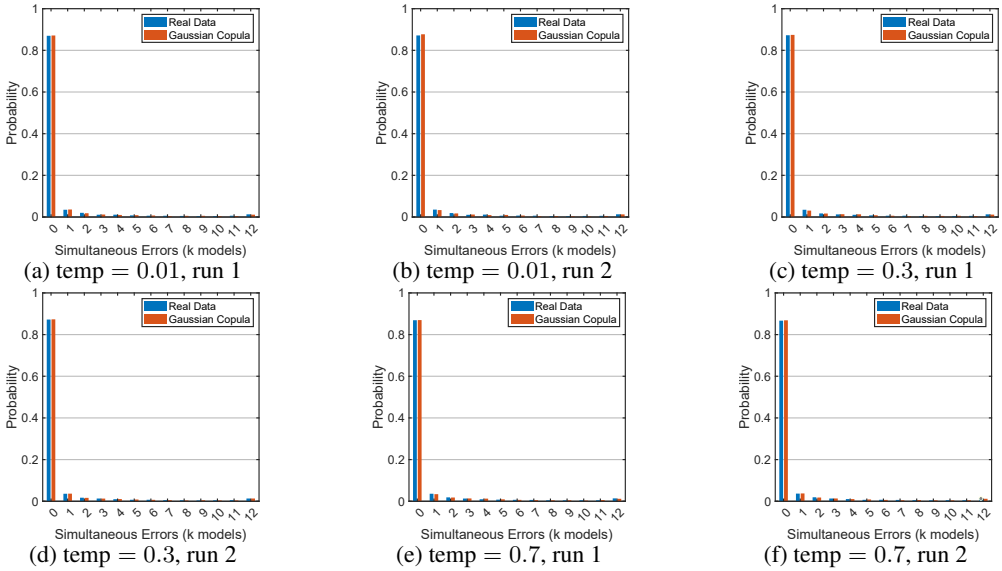
Figure 16: IMDB movie reviews dataset results using **Weighted Ensemble (WE)** aggregation across temperatures and runs.



(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 17: 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 18: Gaussian-copula validation on IMDB: empirical vs. copula-predicted distribution of simultaneous errors across temperatures and runs.

K Additional Experimental Details for the AG News Classification Dataset

In this section, we provide more context on the AG News Classification experiments.

Query Format for AG News Dataset

System: Classify the following news article as WORLD, SPORT, BUSINESS, or SCIENCE/TECHNOLOGY. Respond with only the chosen category. If the article is ambiguous, select the closest matching category.

User: `<article_text>`

We accessed all language models through OpenRouter to standardize the inference process. An actual article from the AG News test split is illustrated below with the ground truth "BUSINESS":

Example Query for AG News Dataset

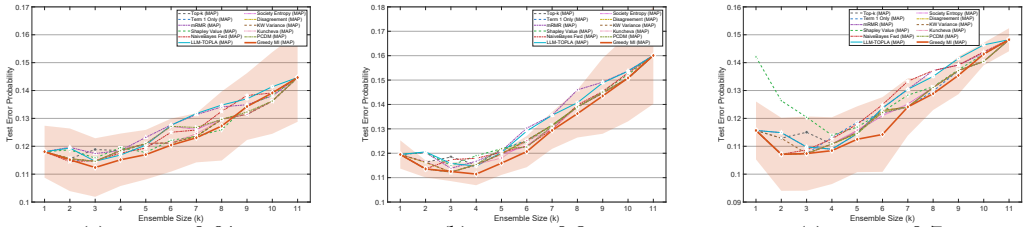
System: Classify the following news article as WORLD, SPORT, BUSINESS, or SCIENCE/TECHNOLOGY. Respond with only the chosen category. If the article is ambiguous, select the closest matching category.

User: Stocks Higher on Drop in Jobless Claims A sharp drop in initial unemployment claims and bullish forecasts from Nokia and Texas Instruments sent stocks higher in early trading Thursday.

LLM: BUSINESS

K.1 Experiments across Temperature Settings (the MAP Aggregation)

For AG News, we also report the experiments for all temperature settings and runs with 10 random splits for each. Figure 19 reports per-condition curves on AG News under the MAP aggregation for each baseline, with one panel for average of each temperature–(10-split) pair. Table 16 aggregates the same experiments over all 30 evaluations (3 temperatures $\times$ 10 random 90/10 splits), reporting mean $\pm$ std test error at each ensemble size k.



(a) temp $= 0.01$ (b) temp $= 0.3$ (c) temp $= 0.7$

Figure 19: Experiments across temperatures and runs (**the MAP aggregation**) on AG News. Each panel corresponds to one temperature setting with average of 10 splits per run. Shaded region represents the standard deviation.

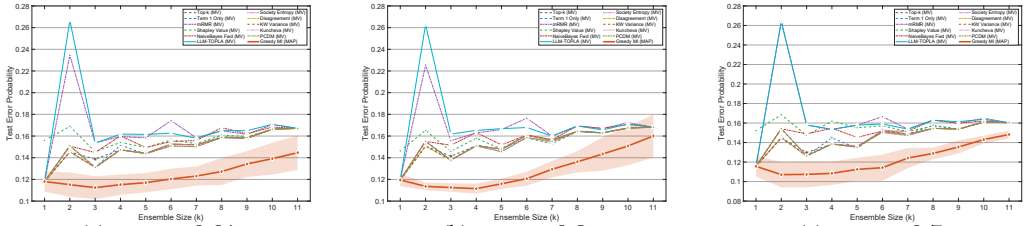
K.2 Experiments with Alternative Aggregation Rules

In this section, to accommodate the multi-class nature of this task, we evaluate the baseline ensembles using our formalized Multi-Class Majority Voting (Algorithm 11) and Multi-Class Weighted Majority Voting (Algorithm 12) strategies. Visualizations from Figure 20 and results in Table 17 reveal that standard Majority Voting (MV) causes severe instability for baselines at small budgets. At $k = 2$, mRMR and LLM-TOPLA error rates spike to 0.241 and 0.265, respectively, while Greedy MI (MAP) achieves an error of 0.112. This performance gap persists through the mid-budget regime; for instance, at $k = 4$, Top-k yields 0.146 and LLM-TOPLA 0.160, compared to 0.112 for Greedy MI.

Applying Weighted Majority Voting (WV), as shown in Figure 21 and Table 18, mitigates some of this instability. However, Greedy MI (MAP) still maintains a distinct advantage in the critical moderate-budget range. For example, at $k = 4$, mRMR reaches 0.134 and at $k = 6$, Top-k reaches 0.128, whereas Greedy MI achieves 0.112 and 0.118, respectively.

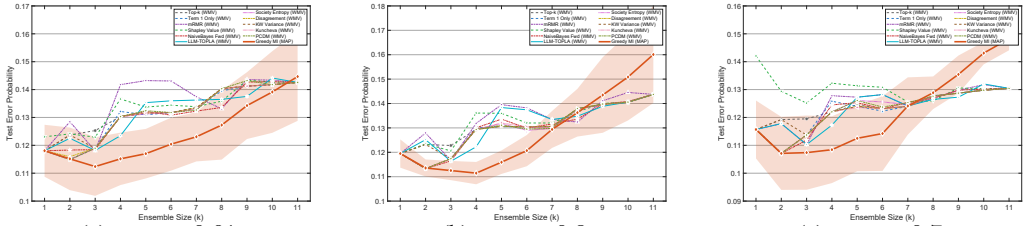
At larger budgets ($k \geq 10$), powerful aggregation rules like WV enable simpler baselines to close the gap and eventually surpass Greedy MI (e.g., Top-k reaches 0.137 at $k = 10$ versus Greedy MI's 0.144). Overall, these results demonstrate that while WV can compensate for poor selection at large k, Greedy MI remains the superior choice in the practical, budget-constrained regime ($k \approx 2$–7) where selection quality matters most.

The five diversity-based baselines follow a similar pattern: at $k = 4$, all five yield an error of 0.146 under MV and 0.127 under WV, compared to 0.112 for Greedy MI (MAP). Under WV, several diversity baselines match Greedy MI at $k = 2$ (0.112), while all five outperform it at $k = 10$ (0.137 versus 0.144), showing that their relative performance depends on the ensemble budget.



(a) temp $= 0.01$ (b) temp $= 0.3$ (c) temp $= 0.7$

Figure 20: AG News Classification dataset results using **Majority Vote (MV)** aggregation across temperatures and runs. Shaded region represents the standard deviation.

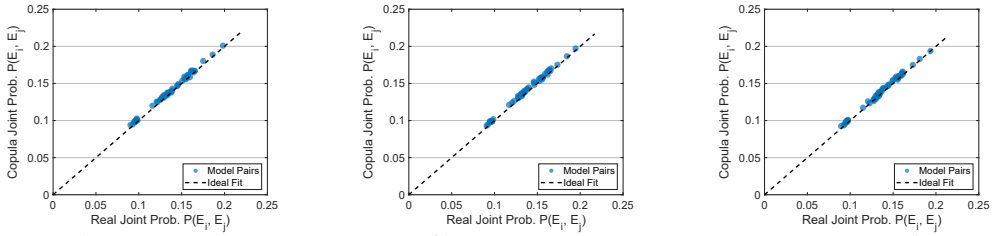


(a) temp $= 0.01$ (b) temp $= 0.3$ (c) temp $= 0.7$

Figure 21: AG News Classification dataset results using **Weighted Ensemble (WE)** aggregation across temperatures and runs.

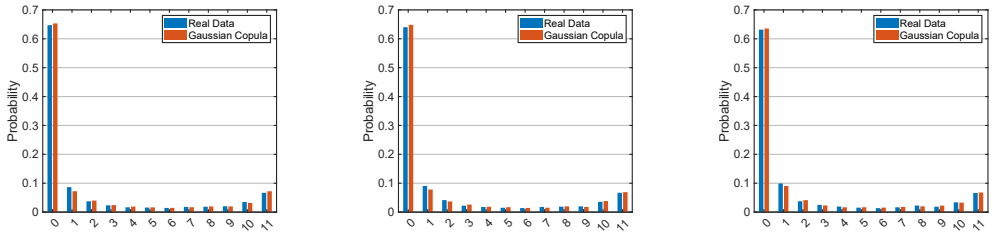
K.3 Copula Validation across Temperature Settings

Figure 22 shows the pairwise fit remains tight along the diagonal across all settings. Figure 23 reveals that the error distribution is heavily concentrated at $k = 0$ and $k = 11$ with the copula accurately capturing this structure.



(a) temp $= 0.01$ (b) temp $= 0.3$ (c) temp $= 0.7$

Figure 22: Gaussian-copula validation on AG News Classification dataset (scatter plots): pairwise structural fit diagnostics across temperatures.



(a) temp $= 0.01$ (b) temp $= 0.3$ (c) temp $= 0.7$

Figure 23: Gaussian-copula validation on AG News Classification dataset (simultaneous-error / tail plots): empirical vs copula-predicted probability of k-way simultaneous errors across temperatures and runs.

L Supplementary Tables

Table 4: Test error probability on MEDMCQA with MAP aggregation for all methods. Results averaged over 30 evaluations (3 temperatures $\times$ 2 runs $\times$ 5 random splits). Bold indicates the lowest displayed mean error for each k, including ties.

k	Top-k	Term 1	mRMR	TMC-Shapley	LogReg	LLM-TOPLA	Society Entropy	Disagreement	KW Variance	Kuncheva	PCDM	Greedy MI (Ours)
1	$\mathbf{0.171 \pm 0.008}$	$\mathbf{0.171 \pm 0.008}$	$\mathbf{0.171 \pm 0.008}$	$\mathbf{0.171 \pm 0.008}$	$\mathbf{0.171 \pm 0.008}$	$\mathbf{0.171 \pm 0.008}$	$\mathbf{0.171 \pm 0.008}$	$\mathbf{0.171 \pm 0.008}$	$\mathbf{0.171 \pm 0.008}$	$\mathbf{0.171 \pm 0.008}$	$\mathbf{0.171 \pm 0.008}$	$\mathbf{0.171 \pm 0.008}$
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	0.172 ± 0.008	0.172 ± 0.008	0.172 ± 0.008	$\mathbf{0.171 \pm 0.008}$	$\mathbf{0.171 \pm 0.008}$	$\mathbf{0.171 \pm 0.008}$
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	0.166 ± 0.009	0.167 ± 0.009	0.174 ± 0.008	$\mathbf{0.164 \pm 0.009}$	$\mathbf{0.164 \pm 0.009}$	$\mathbf{0.164 \pm 0.008}$
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	0.166 ± 0.008	0.166 ± 0.009	0.167 ± 0.009	0.166 ± 0.009	0.166 ± 0.009	$\mathbf{0.163 \pm 0.008}$
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	0.170 ± 0.008	0.170 ± 0.008	0.170 ± 0.008	0.169 ± 0.009	0.170 ± 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	0.171 ± 0.007	0.171 ± 0.007	0.171 ± 0.007	0.171 ± 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	0.170 ± 0.007	0.170 ± 0.007	0.170 ± 0.007	0.172 ± 0.008	0.171 ± 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	0.172 ± 0.007	0.172 ± 0.007	0.172 ± 0.007	0.171 ± 0.008	0.172 ± 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 ± 0.007	0.174 ± 0.007	0.174 ± 0.007	0.174 ± 0.008	0.174 ± 0.008	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	$\mathbf{0.177 \pm 0.008}$	$\mathbf{0.177 \pm 0.008}$	$\mathbf{0.177 \pm 0.008}$	$\mathbf{0.177 \pm 0.007}$	$\mathbf{0.177 \pm 0.007}$	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	0.182 ± 0.009	0.182 ± 0.008	0.182 ± 0.008	$\mathbf{0.181 \pm 0.008}$	$\mathbf{0.181 \pm 0.008}$	0.183 ± 0.008
12	$\mathbf{0.188 \pm 0.009}$	$\mathbf{0.188 \pm 0.009}$	$\mathbf{0.188 \pm 0.009}$	$\mathbf{0.188 \pm 0.009}$	$\mathbf{0.188 \pm 0.009}$	$\mathbf{0.188 \pm 0.009}$	$\mathbf{0.188 \pm 0.009}$	$\mathbf{0.188 \pm 0.009}$	$\mathbf{0.188 \pm 0.009}$	$\mathbf{0.188 \pm 0.009}$	$\mathbf{0.188 \pm 0.009}$	$\mathbf{0.188 \pm 0.009}$

Table 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 lowest displayed mean error for each k, including ties.

k	Top-k (MV)	Term 1 (MV)	mRMR (MV)	TMC-Shapley (MV)	LogReg (MV)	LLM-TOPLA (MV)	Society Entropy (MV)	Disagreement (MV)	KW Variance (MV)	Kuncheva (MV)	PCDM (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	$\mathbf{0.171 \pm 0.008}$	$\mathbf{0.171 \pm 0.008}$	$\mathbf{0.171 \pm 0.008}$	$\mathbf{0.171 \pm 0.008}$	$\mathbf{0.171 \pm 0.008}$	$\mathbf{0.171 \pm 0.008}$	$\mathbf{0.171 \pm 0.008}$	$\mathbf{0.171 \pm 0.008}$
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	0.181 ± 0.009	0.182 ± 0.008	0.181 ± 0.009	0.182 ± 0.008	0.182 ± 0.008	$\mathbf{0.171 \pm 0.008}$
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	0.165 ± 0.009	0.166 ± 0.009	0.172 ± 0.008	$\mathbf{0.164 \pm 0.009}$	$\mathbf{0.164 \pm 0.009}$	$\mathbf{0.164 \pm 0.008}$
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	0.171 ± 0.008	0.171 ± 0.008	0.174 ± 0.009	0.171 ± 0.008	0.171 ± 0.008	$\mathbf{0.163 \pm 0.008}$
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	0.170 ± 0.007	0.170 ± 0.008	0.170 ± 0.008	0.171 ± 0.008	0.171 ± 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	0.170 ± 0.007	0.170 ± 0.007	0.170 ± 0.007	0.171 ± 0.007	0.170 ± 0.007	$\mathbf{0.164 \pm 0.008}$
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	0.172 ± 0.009	0.172 ± 0.009	0.172 ± 0.009	0.171 ± 0.008	0.172 ± 0.008	$\mathbf{0.167 \pm 0.008}$
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	0.173 ± 0.008	0.173 ± 0.008	0.173 ± 0.008	0.172 ± 0.007	0.172 ± 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	0.181 ± 0.008	0.181 ± 0.008	0.181 ± 0.008	0.180 ± 0.009	0.180 ± 0.009	$\mathbf{0.174 \pm 0.008}$
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	0.183 ± 0.008	0.183 ± 0.008	0.183 ± 0.008	$\mathbf{0.180 \pm 0.007}$	$\mathbf{0.180 \pm 0.007}$	0.181 ± 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	0.191 ± 0.008	0.192 ± 0.008	0.193 ± 0.008	0.190 ± 0.008	0.190 ± 0.008	$\mathbf{0.183 \pm 0.008}$
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	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 lowest displayed mean error for each k, including ties.

k	Top-k (WV)	Term 1 (WV)	mRMR (WV)	TMC-Shapley (WV)	LogReg (WV)	LLM-TOPLA (WV)	Society Entropy (WV)	Disagreement (WV)	KW Variance (WV)	Kuncheva (WV)	PCDM (WV)	Greedy MI (MAP)
1	$\mathbf{0.171 \pm 0.008}$	$\mathbf{0.171 \pm 0.008}$	$\mathbf{0.171 \pm 0.008}$	$\mathbf{0.171 \pm 0.008}$	$\mathbf{0.171 \pm 0.008}$	$\mathbf{0.171 \pm 0.008}$	$\mathbf{0.171 \pm 0.008}$	$\mathbf{0.171 \pm 0.008}$	$\mathbf{0.171 \pm 0.008}$	$\mathbf{0.171 \pm 0.008}$	$\mathbf{0.171 \pm 0.008}$	$\mathbf{0.171 \pm 0.008}$
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	$\mathbf{0.171 \pm 0.008}$	$\mathbf{0.171 \pm 0.008}$	$\mathbf{0.171 \pm 0.008}$	$\mathbf{0.171 \pm 0.008}$	$\mathbf{0.171 \pm 0.008}$	$\mathbf{0.171 \pm 0.008}$
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	0.165 ± 0.009	0.166 ± 0.009	0.172 ± 0.008	$\mathbf{0.164 \pm 0.009}$	$\mathbf{0.164 \pm 0.009}$	$\mathbf{0.164 \pm 0.008}$
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	0.172 ± 0.008	0.172 ± 0.008	0.172 ± 0.008	0.166 ± 0.009	0.168 ± 0.009	$\mathbf{0.163 \pm 0.008}$
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	0.170 ± 0.007	0.170 ± 0.008	0.170 ± 0.008	0.171 ± 0.008	0.171 ± 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	0.170 ± 0.009	0.170 ± 0.009	0.170 ± 0.009	0.170 ± 0.008	0.170 ± 0.008	$\mathbf{0.164 \pm 0.008}$
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	0.172 ± 0.009	0.172 ± 0.009	0.172 ± 0.009	0.171 ± 0.008	0.172 ± 0.008	$\mathbf{0.167 \pm 0.008}$
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	0.172 ± 0.009	0.172 ± 0.009	0.172 ± 0.009	$\mathbf{0.169 \pm 0.008}$	$\mathbf{0.169 \pm 0.008}$	0.170 ± 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	0.181 ± 0.008	0.181 ± 0.008	0.181 ± 0.008	0.180 ± 0.009	0.180 ± 0.009	$\mathbf{0.174 \pm 0.008}$
10	0.181 ± 0.009	0.181 ± 0.009	0.185 ± 0.007	0.178 ± 0.009	0.188 ± 0.009	0.180 ± 0.007	0.181 ± 0.009	0.181 ± 0.009	0.181 ± 0.009	$\mathbf{0.176 \pm 0.008}$	$\mathbf{0.176 \pm 0.008}$	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	0.191 ± 0.008	0.192 ± 0.008	0.193 ± 0.008	0.190 ± 0.008	0.190 ± 0.008	$\mathbf{0.183 \pm 0.008}$
12	$\mathbf{0.188 \pm 0.009}$	$\mathbf{0.188 \pm 0.009}$	$\mathbf{0.188 \pm 0.009}$	$\mathbf{0.188 \pm 0.009}$	$\mathbf{0.188 \pm 0.009}$	$\mathbf{0.188 \pm 0.009}$	$\mathbf{0.188 \pm 0.009}$	$\mathbf{0.188 \pm 0.009}$	$\mathbf{0.188 \pm 0.009}$	$\mathbf{0.188 \pm 0.009}$	$\mathbf{0.188 \pm 0.009}$	$\mathbf{0.188 \pm 0.009}$

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%	—	$\bar{\rho} = 0.55$

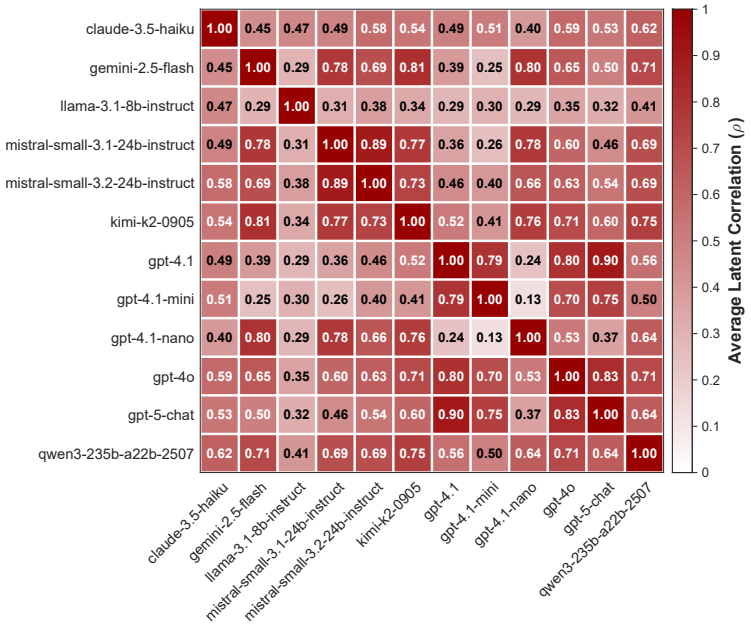


Figure 24: Averaged corr. matrix for MEDMCQA for 3 temperature settings and 2 runs.

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 random splits). Bold indicates the lowest displayed mean error for each k, including ties.

k	Top-k	Term 1	mRMR	Shapley Value	LogReg	LLM-TOPLA	Society Entropy	Disagreement	KW Variance	Kuncheva	PCDM	Greedy MI
1	$\mathbf{0.151 \pm 0.007}$	$\mathbf{0.151 \pm 0.007}$	$\mathbf{0.151 \pm 0.007}$	0.154 ± 0.009	$\mathbf{0.151 \pm 0.007}$	$\mathbf{0.151 \pm 0.007}$	$\mathbf{0.151 \pm 0.007}$	$\mathbf{0.151 \pm 0.007}$	$\mathbf{0.151 \pm 0.007}$	$\mathbf{0.151 \pm 0.007}$	$\mathbf{0.151 \pm 0.007}$	$\mathbf{0.151 \pm 0.007}$
2	0.152 ± 0.007	0.152 ± 0.007	$\mathbf{0.151 \pm 0.007}$	0.153 ± 0.008	$\mathbf{0.151 \pm 0.007}$	$\mathbf{0.151 \pm 0.007}$	$\mathbf{0.151 \pm 0.008}$	$\mathbf{0.151 \pm 0.008}$	0.152 ± 0.008	$\mathbf{0.151 \pm 0.008}$	$\mathbf{0.151 \pm 0.008}$	$\mathbf{0.151 \pm 0.008}$
3	0.153 ± 0.007	0.153 ± 0.007	0.151 ± 0.007	0.154 ± 0.008	0.150 ± 0.006	0.152 ± 0.008	0.152 ± 0.007	0.153 ± 0.007	0.153 ± 0.007	0.153 ± 0.007	0.153 ± 0.007	$\mathbf{0.148 \pm 0.008}$
4	0.151 ± 0.005	0.150 ± 0.005	0.151 ± 0.007	0.149 ± 0.006	0.147 ± 0.007	0.151 ± 0.007	0.151 ± 0.005	0.151 ± 0.005	0.151 ± 0.005	0.151 ± 0.005	0.151 ± 0.005	$\mathbf{0.144 \pm 0.007}$
5	0.151 ± 0.006	0.150 ± 0.006	0.150 ± 0.007	0.149 ± 0.006	0.148 ± 0.006	0.152 ± 0.007	0.151 ± 0.006	0.151 ± 0.006	0.151 ± 0.006	0.151 ± 0.006	0.151 ± 0.006	$\mathbf{0.141 \pm 0.008}$
6	0.149 ± 0.007	0.149 ± 0.007	0.149 ± 0.007	0.150 ± 0.008	0.149 ± 0.006	0.150 ± 0.006	0.149 ± 0.007	0.149 ± 0.007	0.149 ± 0.007	0.149 ± 0.007	0.149 ± 0.007	$\mathbf{0.142 \pm 0.008}$
7	0.150 ± 0.007	0.150 ± 0.007	0.151 ± 0.007	0.149 ± 0.007	0.150 ± 0.007	0.149 ± 0.007	0.150 ± 0.007	0.150 ± 0.007	0.150 ± 0.007	0.150 ± 0.007	0.150 ± 0.007	$\mathbf{0.144 \pm 0.006}$
8	0.152 ± 0.007	0.152 ± 0.007	0.151 ± 0.006	0.151 ± 0.006	0.150 ± 0.006	0.151 ± 0.007	0.152 ± 0.007	0.152 ± 0.007	0.152 ± 0.007	0.152 ± 0.007	0.152 ± 0.007	$\mathbf{0.149 \pm 0.006}$
9	0.153 ± 0.007	$\mathbf{0.152 \pm 0.007}$	0.153 ± 0.007	0.153 ± 0.007	0.154 ± 0.007	$\mathbf{0.152 \pm 0.007}$	0.154 ± 0.007	0.153 ± 0.007	0.153 ± 0.007	0.153 ± 0.007	0.154 ± 0.007	$\mathbf{0.152 \pm 0.007}$
10	$\mathbf{0.156 \pm 0.007}$	$\mathbf{0.156 \pm 0.007}$	$\mathbf{0.156 \pm 0.007}$	0.157 ± 0.007	$\mathbf{0.156 \pm 0.008}$	$\mathbf{0.156 \pm 0.006}$	$\mathbf{0.156 \pm 0.007}$	$\mathbf{0.156 \pm 0.007}$	$\mathbf{0.156 \pm 0.007}$	$\mathbf{0.156 \pm 0.007}$	$\mathbf{0.156 \pm 0.007}$	0.157 ± 0.007
11	$\mathbf{0.158 \pm 0.007}$	$\mathbf{0.158 \pm 0.007}$	0.160 ± 0.007	0.160 ± 0.008	0.160 ± 0.008	0.161 ± 0.008	$\mathbf{0.158 \pm 0.007}$	$\mathbf{0.158 \pm 0.007}$	$\mathbf{0.158 \pm 0.007}$	0.160 ± 0.007	0.160 ± 0.007	$\mathbf{0.158 \pm 0.007}$
12	$\mathbf{0.164 \pm 0.008}$	$\mathbf{0.164 \pm 0.008}$	0.166 ± 0.006	0.165 ± 0.008	0.166 ± 0.007	0.167 ± 0.007	$\mathbf{0.164 \pm 0.008}$	$\mathbf{0.164 \pm 0.008}$	$\mathbf{0.164 \pm 0.008}$	$\mathbf{0.164 \pm 0.007}$	0.165 ± 0.009	$\mathbf{0.164 \pm 0.008}$
13	$\mathbf{0.171 \pm 0.008}$	$\mathbf{0.171 \pm 0.008}$	$\mathbf{0.171 \pm 0.008}$	$\mathbf{0.171 \pm 0.008}$	$\mathbf{0.171 \pm 0.008}$	$\mathbf{0.171 \pm 0.008}$	$\mathbf{0.171 \pm 0.008}$	$\mathbf{0.171 \pm 0.008}$	$\mathbf{0.171 \pm 0.008}$	$\mathbf{0.171 \pm 0.008}$	$\mathbf{0.171 \pm 0.008}$	$\mathbf{0.171 \pm 0.008}$

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 lowest displayed mean error for each k, including ties.

k	Top-k (MV)	Term 1 (MV)	mRMR (MV)	Shapley Value (MV)	LogReg (MV)	LLM-TOPLA (MV)	Society Entropy (MV)	Disagreement (MV)	KW Variance (MV)	Kuncheva (MV)	PCDM (MV)	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	$\mathbf{0.151 \pm 0.007}$	$\mathbf{0.151 \pm 0.007}$	$\mathbf{0.151 \pm 0.007}$	$\mathbf{0.151 \pm 0.007}$	$\mathbf{0.151 \pm 0.007}$	$\mathbf{0.151 \pm 0.007}$	$\mathbf{0.151 \pm 0.007}$	$\mathbf{0.151 \pm 0.007}$
2	0.157 ± 0.006	0.157 ± 0.006	0.247 ± 0.022	0.158 ± 0.010	$\mathbf{0.151 \pm 0.008}$	0.183 ± 0.049	0.153 ± 0.008	0.153 ± 0.008	0.155 ± 0.008	0.153 ± 0.008	0.153 ± 0.008	$\mathbf{0.151 \pm 0.008}$
3	0.151 ± 0.007	0.151 ± 0.007	0.219 ± 0.026	0.153 ± 0.007	0.158 ± 0.013	0.167 ± 0.011	0.150 ± 0.007	0.151 ± 0.007	0.151 ± 0.007	0.151 ± 0.007	0.151 ± 0.007	$\mathbf{0.148 \pm 0.008}$
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	0.151 ± 0.006	0.151 ± 0.006	0.151 ± 0.006	0.151 ± 0.006	0.151 ± 0.006	$\mathbf{0.144 \pm 0.007}$
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	0.148 ± 0.006	0.148 ± 0.006	0.148 ± 0.006	0.149 ± 0.006	0.148 ± 0.006	$\mathbf{0.141 \pm 0.008}$
6	0.145 ± 0.006	0.145 ± 0.006	0.176 ± 0.011	0.147 ± 0.007	0.151 ± 0.008	0.156 ± 0.007	0.145 ± 0.006	0.145 ± 0.006	0.145 ± 0.006	0.145 ± 0.006	0.145 ± 0.006	$\mathbf{0.142 \pm 0.008}$
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	0.148 ± 0.006	0.148 ± 0.006	0.148 ± 0.006	0.148 ± 0.006	0.148 ± 0.006	$\mathbf{0.144 \pm 0.006}$
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	$\mathbf{0.149 \pm 0.005}$	$\mathbf{0.149 \pm 0.005}$	$\mathbf{0.149 \pm 0.005}$	0.150 ± 0.006	0.150 ± 0.006	$\mathbf{0.149 \pm 0.006}$
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	0.154 ± 0.005	0.154 ± 0.005	0.154 ± 0.005	0.155 ± 0.005	0.155 ± 0.005	$\mathbf{0.152 \pm 0.007}$
10	0.155 ± 0.006	0.155 ± 0.005	0.167 ± 0.008	0.155 ± 0.006	0.162 ± 0.009	0.158 ± 0.007	0.155 ± 0.006	0.155 ± 0.006	0.155 ± 0.006	$\mathbf{0.154 \pm 0.005}$	0.155 ± 0.006	0.157 ± 0.007
11	0.165 ± 0.007	0.165 ± 0.007	0.172 ± 0.009	0.165 ± 0.008	0.170 ± 0.010	0.165 ± 0.008	0.165 ± 0.007	0.165 ± 0.007	0.165 ± 0.007	0.164 ± 0.007	0.164 ± 0.007	$\mathbf{0.158 \pm 0.007}$
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	0.164 ± 0.008	0.164 ± 0.008	0.164 ± 0.008	0.164 ± 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	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 lowest displayed mean error for each k, including ties.

k	Top-k (WV)	Term 1 (WV)	mRMR (WV)	Shapley Value (WV)	LogReg (WV)	LLM-TOPLA (WV)	Society Entropy (WV)	Disagreement (WV)	KW Variance (WV)	Kuncheva (WV)	PCDM (WV)	Greedy MI (MAP)
1	$\mathbf{0.151 \pm 0.007}$	$\mathbf{0.151 \pm 0.007}$	$\mathbf{0.151 \pm 0.007}$	0.153 ± 0.008	$\mathbf{0.151 \pm 0.007}$	$\mathbf{0.151 \pm 0.007}$	$\mathbf{0.151 \pm 0.007}$	$\mathbf{0.151 \pm 0.007}$	$\mathbf{0.151 \pm 0.007}$	$\mathbf{0.151 \pm 0.007}$	$\mathbf{0.151 \pm 0.007}$	$\mathbf{0.151 \pm 0.007}$
2	$\mathbf{0.151 \pm 0.007}$	$\mathbf{0.151 \pm 0.007}$	$\mathbf{0.151 \pm 0.007}$	$\mathbf{0.151 \pm 0.007}$	$\mathbf{0.151 \pm 0.007}$	$\mathbf{0.151 \pm 0.007}$	$\mathbf{0.151 \pm 0.007}$	$\mathbf{0.151 \pm 0.007}$	$\mathbf{0.151 \pm 0.007}$	$\mathbf{0.151 \pm 0.007}$	$\mathbf{0.151 \pm 0.007}$	$\mathbf{0.151 \pm 0.008}$
3	0.151 ± 0.007	0.151 ± 0.007	0.219 ± 0.026	0.153 ± 0.007	0.158 ± 0.013	0.167 ± 0.011	0.150 ± 0.007	0.151 ± 0.007	0.151 ± 0.007	0.151 ± 0.007	0.151 ± 0.007	$\mathbf{0.148 \pm 0.008}$
4	0.149 ± 0.007	0.149 ± 0.007	0.175 ± 0.009	0.149 ± 0.006	0.151 ± 0.008	0.155 ± 0.009	0.150 ± 0.007	0.149 ± 0.007	0.149 ± 0.007	0.150 ± 0.007	0.150 ± 0.007	$\mathbf{0.144 \pm 0.007}$
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	0.148 ± 0.006	0.148 ± 0.006	0.148 ± 0.006	0.149 ± 0.006	0.148 ± 0.006	$\mathbf{0.141 \pm 0.008}$
6	0.147 ± 0.006	0.147 ± 0.006	0.170 ± 0.012	0.148 ± 0.007	0.151 ± 0.007	0.155 ± 0.007	0.147 ± 0.006	0.147 ± 0.006	0.147 ± 0.006	0.147 ± 0.006	0.147 ± 0.006	$\mathbf{0.142 \pm 0.008}$
7	0.148 ± 0.006	0.149 ± 0.006	0.184 ± 0.014	0.148 ± 0.006	0.159 ± 0.012	0.158 ± 0.009	0.148 ± 0.006	0.148 ± 0.006	0.148 ± 0.006	0.148 ± 0.006	0.148 ± 0.006	$\mathbf{0.144 \pm 0.006}$
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	$\mathbf{0.148 \pm 0.006}$	$\mathbf{0.148 \pm 0.006}$	$\mathbf{0.148 \pm 0.006}$	0.149 ± 0.006	0.149 ± 0.006	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	0.154 ± 0.005	0.154 ± 0.005	0.154 ± 0.005	0.155 ± 0.005	0.155 ± 0.005	$\mathbf{0.152 \pm 0.007}$
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	$\mathbf{0.153 \pm 0.006}$	$\mathbf{0.153 \pm 0.006}$	$\mathbf{0.153 \pm 0.006}$	$\mathbf{0.153 \pm 0.006}$	$\mathbf{0.153 \pm 0.006}$	0.157 ± 0.007
11	0.165 ± 0.007	0.165 ± 0.007	0.171 ± 0.009	0.164 ± 0.007	0.170 ± 0.010	0.165 ± 0.008	0.165 ± 0.007	0.165 ± 0.007	0.165 ± 0.007	0.164 ± 0.007	0.164 ± 0.007	$\mathbf{0.158 \pm 0.007}$
12	0.163 ± 0.007	0.163 ± 0.007	0.162 ± 0.008	0.162 ± 0.008	0.164 ± 0.007	$\mathbf{0.160 \pm 0.008}$	0.163 ± 0.007	0.163 ± 0.007	0.163 ± 0.007	0.163 ± 0.007	0.163 ± 0.007	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	0.174 ± 0.009	0.174 ± 0.009	0.174 ± 0.009	0.174 ± 0.009	0.174 ± 0.009	$\mathbf{0.171 \pm 0.008}$

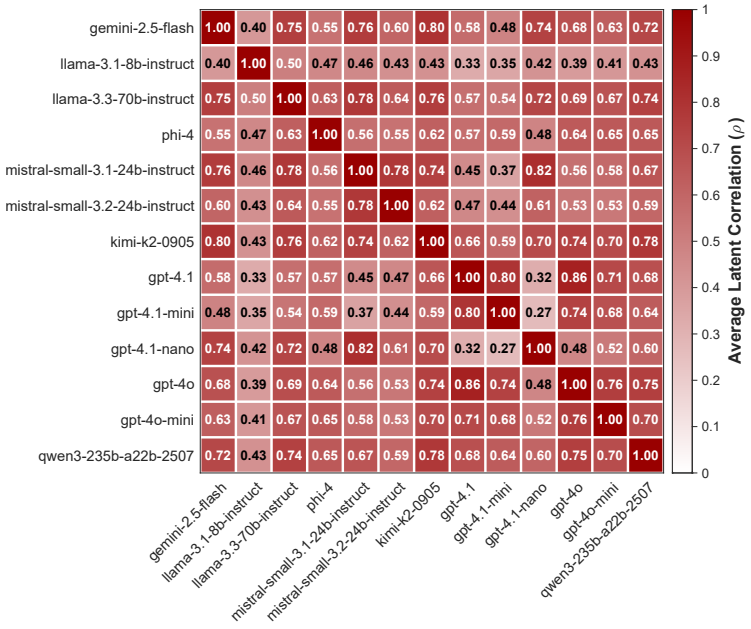


Figure 25: Averaged corr. matrix for MMLU for 3 temperature settings and 2 runs.

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 random splits). Bold indicates the lowest displayed mean error for each k, including ties.

k	Top-k	Term 1	mRMR	TMC-Shapley	LogReg	LLM-TOPLA	Society Entropy	Disagreement	KW Variance	Kuncheva	PCDM	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	$\mathbf{0.035 \pm 0.004}$	$\mathbf{0.035 \pm 0.004}$	$\mathbf{0.035 \pm 0.004}$	$\mathbf{0.035 \pm 0.004}$	$\mathbf{0.035 \pm 0.004}$	$\mathbf{0.035 \pm 0.004}$	$\mathbf{0.035 \pm 0.004}$	$\mathbf{0.035 \pm 0.004}$
2	0.036 ± 0.004	0.036 ± 0.004	$\mathbf{0.035 \pm 0.004}$	0.036 ± 0.004	$\mathbf{0.035 \pm 0.004}$	$\mathbf{0.035 \pm 0.004}$	0.036 ± 0.004	0.036 ± 0.004	0.036 ± 0.004	0.036 ± 0.004	0.036 ± 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	$\mathbf{0.034 \pm 0.004}$	$\mathbf{0.034 \pm 0.004}$	$\mathbf{0.034 \pm 0.004}$	$\mathbf{0.034 \pm 0.004}$	$\mathbf{0.034 \pm 0.004}$	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	0.036 ± 0.004	0.036 ± 0.004	0.036 ± 0.004	0.036 ± 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	0.038 ± 0.004	0.038 ± 0.004	0.038 ± 0.004	0.037 ± 0.005	0.037 ± 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	0.039 ± 0.004	0.039 ± 0.004	0.039 ± 0.004	0.039 ± 0.004	0.038 ± 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	0.039 ± 0.004	0.040 ± 0.004	0.040 ± 0.004	0.039 ± 0.004	0.038 ± 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	0.040 ± 0.004	0.040 ± 0.004	0.040 ± 0.004	0.040 ± 0.004	0.040 ± 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	0.039 ± 0.004	0.039 ± 0.004	0.039 ± 0.004	0.039 ± 0.004	0.039 ± 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	$\mathbf{0.038 \pm 0.004}$	0.040 ± 0.004	0.040 ± 0.004	0.040 ± 0.004	0.040 ± 0.004	0.040 ± 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	$\mathbf{0.040 \pm 0.004}$	$\mathbf{0.040 \pm 0.004}$	0.041 ± 0.004	0.041 ± 0.004	0.041 ± 0.004	0.041 ± 0.004	0.041 ± 0.004	$\mathbf{0.040 \pm 0.004}$
12	$\mathbf{0.041 \pm 0.005}$	$\mathbf{0.041 \pm 0.005}$	$\mathbf{0.041 \pm 0.005}$	$\mathbf{0.041 \pm 0.005}$	$\mathbf{0.041 \pm 0.005}$	$\mathbf{0.041 \pm 0.005}$	$\mathbf{0.041 \pm 0.005}$	$\mathbf{0.041 \pm 0.005}$	$\mathbf{0.041 \pm 0.005}$	$\mathbf{0.041 \pm 0.005}$	$\mathbf{0.041 \pm 0.005}$	$\mathbf{0.041 \pm 0.005}$

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 lowest displayed mean error for each k, including ties.

k	Top-k (MV)	Term 1 (MV)	mRMR (MV)	TMC-Shapley (MV)	LogReg (MV)	LLM-TOPLA (MV)	Society Entropy (MV)	Disagreement (MV)	KW Variance (MV)	Kuncheva (MV)	PCDM (MV)	Greedy MI (MAP)
1	$\mathbf{0.035 \pm 0.004}$	$\mathbf{0.035 \pm 0.004}$	$\mathbf{0.035 \pm 0.004}$	0.038 ± 0.005	$\mathbf{0.035 \pm 0.004}$	$\mathbf{0.035 \pm 0.004}$	$\mathbf{0.035 \pm 0.004}$	$\mathbf{0.035 \pm 0.004}$	$\mathbf{0.035 \pm 0.004}$	$\mathbf{0.035 \pm 0.004}$	$\mathbf{0.035 \pm 0.004}$	$\mathbf{0.035 \pm 0.004}$
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	$\mathbf{0.035 \pm 0.004}$	$\mathbf{0.035 \pm 0.004}$	$\mathbf{0.035 \pm 0.004}$	$\mathbf{0.035 \pm 0.004}$	$\mathbf{0.035 \pm 0.004}$	$\mathbf{0.035 \pm 0.004}$
3	$\mathbf{0.034 \pm 0.004}$	$\mathbf{0.034 \pm 0.004}$	0.039 ± 0.004	0.035 ± 0.004	0.036 ± 0.005	0.041 ± 0.004	$\mathbf{0.034 \pm 0.004}$	$\mathbf{0.034 \pm 0.004}$	$\mathbf{0.034 \pm 0.004}$	0.035 ± 0.004	0.035 ± 0.004	0.036 ± 0.004
4	0.036 ± 0.004	0.036 ± 0.004	0.037 ± 0.004	0.036 ± 0.004	0.036 ± 0.004	0.037 ± 0.004	0.035 ± 0.004	0.036 ± 0.004	0.036 ± 0.004	$\mathbf{0.034 \pm 0.004}$	$\mathbf{0.034 \pm 0.004}$	$\mathbf{0.034 \pm 0.004}$
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	0.036 ± 0.004	0.036 ± 0.004	0.036 ± 0.004	0.036 ± 0.004	0.036 ± 0.004	$\mathbf{0.033 \pm 0.004}$
6	0.036 ± 0.004	0.036 ± 0.004	0.036 ± 0.004	0.036 ± 0.004	0.038 ± 0.004	0.037 ± 0.004	0.036 ± 0.004	0.036 ± 0.004	0.036 ± 0.004	0.036 ± 0.004	0.036 ± 0.004	$\mathbf{0.033 \pm 0.004}$
7	0.036 ± 0.004	0.036 ± 0.004	0.038 ± 0.004	0.036 ± 0.004	0.040 ± 0.005	0.040 ± 0.004	0.036 ± 0.004	0.036 ± 0.004	0.036 ± 0.004	0.036 ± 0.004	0.036 ± 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	0.036 ± 0.004	0.036 ± 0.004	0.036 ± 0.004	0.036 ± 0.004	0.036 ± 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	$\mathbf{0.037 \pm 0.004}$	$\mathbf{0.037 \pm 0.004}$	$\mathbf{0.037 \pm 0.004}$	$\mathbf{0.037 \pm 0.004}$	$\mathbf{0.037 \pm 0.004}$	$\mathbf{0.037 \pm 0.004}$
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	0.038 ± 0.004	0.038 ± 0.004	0.038 ± 0.004	0.038 ± 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	$\mathbf{0.040 \pm 0.004}$	$\mathbf{0.040 \pm 0.004}$	$\mathbf{0.040 \pm 0.004}$	$\mathbf{0.040 \pm 0.004}$	$\mathbf{0.040 \pm 0.004}$	$\mathbf{0.040 \pm 0.004}$
12	$\mathbf{0.040 \pm 0.004}$	$\mathbf{0.040 \pm 0.004}$	$\mathbf{0.040 \pm 0.004}$	$\mathbf{0.040 \pm 0.004}$	$\mathbf{0.040 \pm 0.004}$	$\mathbf{0.040 \pm 0.004}$	$\mathbf{0.040 \pm 0.004}$	$\mathbf{0.040 \pm 0.004}$	$\mathbf{0.040 \pm 0.004}$	$\mathbf{0.040 \pm 0.004}$	$\mathbf{0.040 \pm 0.004}$	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 lowest displayed mean error for each k, including ties.

k	Top-k (WV)	Term 1 (WV)	mRMR (WV)	TMC-Shapley (WV)	LogReg (WV)	LLM-TOPLA (WV)	Society Entropy (WV)	Disagreement (WV)	KW Variance (WV)	Kuncheva (WV)	PCDM (WV)	Greedy MI (MAP)
1	$\mathbf{0.035 \pm 0.004}$	$\mathbf{0.035 \pm 0.004}$	$\mathbf{0.035 \pm 0.004}$	0.038 ± 0.005	$\mathbf{0.035 \pm 0.004}$	$\mathbf{0.035 \pm 0.004}$	$\mathbf{0.035 \pm 0.004}$	$\mathbf{0.035 \pm 0.004}$	$\mathbf{0.035 \pm 0.004}$	$\mathbf{0.035 \pm 0.004}$	$\mathbf{0.035 \pm 0.004}$	$\mathbf{0.035 \pm 0.004}$
2	$\mathbf{0.035 \pm 0.004}$	$\mathbf{0.035 \pm 0.004}$	$\mathbf{0.035 \pm 0.004}$	$\mathbf{0.035 \pm 0.004}$	$\mathbf{0.035 \pm 0.004}$	$\mathbf{0.035 \pm 0.004}$	$\mathbf{0.035 \pm 0.004}$	$\mathbf{0.035 \pm 0.004}$	$\mathbf{0.035 \pm 0.004}$	$\mathbf{0.035 \pm 0.004}$	$\mathbf{0.035 \pm 0.004}$	$\mathbf{0.035 \pm 0.004}$
3	$\mathbf{0.034 \pm 0.004}$	$\mathbf{0.034 \pm 0.004}$	0.039 ± 0.004	0.035 ± 0.004	0.036 ± 0.005	0.041 ± 0.004	$\mathbf{0.034 \pm 0.004}$	$\mathbf{0.034 \pm 0.004}$	$\mathbf{0.034 \pm 0.004}$	0.035 ± 0.004	0.035 ± 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	$\mathbf{0.034 \pm 0.004}$	$\mathbf{0.034 \pm 0.004}$	$\mathbf{0.034 \pm 0.004}$	$\mathbf{0.034 \pm 0.004}$	$\mathbf{0.034 \pm 0.004}$	$\mathbf{0.034 \pm 0.004}$
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	0.036 ± 0.004	0.036 ± 0.004	0.036 ± 0.004	0.036 ± 0.004	0.036 ± 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	0.035 ± 0.004	0.035 ± 0.004	0.035 ± 0.004	0.035 ± 0.004	0.035 ± 0.004	$\mathbf{0.033 \pm 0.004}$
7	0.036 ± 0.004	0.036 ± 0.004	0.038 ± 0.004	0.036 ± 0.004	0.040 ± 0.005	0.040 ± 0.004	0.036 ± 0.004	0.036 ± 0.004	0.036 ± 0.004	0.036 ± 0.004	0.036 ± 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	$\mathbf{0.035 \pm 0.004}$	$\mathbf{0.035 \pm 0.004}$	$\mathbf{0.035 \pm 0.004}$	$\mathbf{0.035 \pm 0.004}$	$\mathbf{0.035 \pm 0.004}$	$\mathbf{0.035 \pm 0.004}$
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	$\mathbf{0.037 \pm 0.004}$	$\mathbf{0.037 \pm 0.004}$	$\mathbf{0.037 \pm 0.004}$	$\mathbf{0.037 \pm 0.004}$	$\mathbf{0.037 \pm 0.004}$	$\mathbf{0.037 \pm 0.004}$
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	$\mathbf{0.037 \pm 0.004}$	$\mathbf{0.037 \pm 0.004}$	$\mathbf{0.037 \pm 0.004}$	$\mathbf{0.037 \pm 0.004}$	$\mathbf{0.037 \pm 0.004}$	0.038 ± 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	$\mathbf{0.040 \pm 0.004}$	$\mathbf{0.040 \pm 0.004}$	$\mathbf{0.040 \pm 0.004}$	$\mathbf{0.040 \pm 0.004}$	$\mathbf{0.040 \pm 0.004}$	$\mathbf{0.040 \pm 0.004}$
12	$\mathbf{0.040 \pm 0.004}$	$\mathbf{0.040 \pm 0.004}$	$\mathbf{0.040 \pm 0.004}$	$\mathbf{0.040 \pm 0.004}$	$\mathbf{0.040 \pm 0.004}$	$\mathbf{0.040 \pm 0.004}$	$\mathbf{0.040 \pm 0.004}$	$\mathbf{0.040 \pm 0.004}$	$\mathbf{0.040 \pm 0.004}$	$\mathbf{0.040 \pm 0.004}$	$\mathbf{0.040 \pm 0.004}$	0.041 ± 0.005

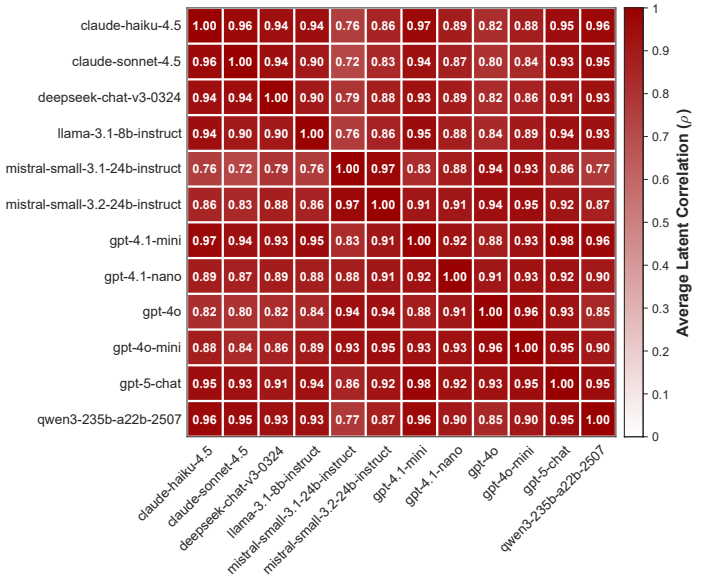


Figure 26: Averaged corr. matrix for IMDB movie reviews for 3 temperature settings and 2 runs.

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	0.840
Average	94.54%	—	$\bar{\rho} = 0.90$

Table 16: Test error probability on AG News with MAP aggregation for all methods. Results averaged over 30 evaluations (3 temperatures $\times$ 10 random 90/10 splits). Bold indicates the lowest displayed mean error for each k, including ties.

k	Top-k	Term 1	mRMR	TMC-Shapley	NaiveBayes FwD	LLM-TOPLA	Society Entropy	Disagreement	KW Variance	Kuncheva	PCDM	Greedy MI
1	$\mathbf{0.118 \pm 0.009}$	$\mathbf{0.118 \pm 0.009}$	$\mathbf{0.118 \pm 0.009}$	0.127 ± 0.019	$\mathbf{0.118 \pm 0.009}$	$\mathbf{0.118 \pm 0.009}$	$\mathbf{0.118 \pm 0.009}$	$\mathbf{0.118 \pm 0.009}$	$\mathbf{0.118 \pm 0.009}$	$\mathbf{0.118 \pm 0.009}$	$\mathbf{0.118 \pm 0.009}$	$\mathbf{0.118 \pm 0.009}$
2	0.115 ± 0.010	$\mathbf{0.112 \pm 0.010}$	0.118 ± 0.009	0.122 ± 0.008	$\mathbf{0.112 \pm 0.010}$	0.118 ± 0.009	0.113 ± 0.011	0.113 ± 0.011	0.115 ± 0.010	$\mathbf{0.112 \pm 0.010}$	$\mathbf{0.112 \pm 0.010}$	$\mathbf{0.112 \pm 0.010}$
3	0.117 ± 0.010	$\mathbf{0.111 \pm 0.011}$	0.114 ± 0.010	0.117 ± 0.007	0.114 ± 0.011	0.113 ± 0.009	$\mathbf{0.111 \pm 0.011}$	$\mathbf{0.111 \pm 0.011}$	$\mathbf{0.111 \pm 0.011}$	$\mathbf{0.111 \pm 0.011}$	$\mathbf{0.111 \pm 0.011}$	$\mathbf{0.111 \pm 0.010}$
4	0.115 ± 0.010	0.116 ± 0.009	0.115 ± 0.011	0.117 ± 0.009	0.116 ± 0.009	0.114 ± 0.011	0.115 ± 0.010	0.115 ± 0.010	0.115 ± 0.010	0.115 ± 0.010	0.115 ± 0.010	$\mathbf{0.112 \pm 0.009}$
5	0.119 ± 0.009	0.120 ± 0.009	0.120 ± 0.010	0.119 ± 0.009	0.119 ± 0.009	0.118 ± 0.011	0.119 ± 0.008	0.118 ± 0.008	0.119 ± 0.009	0.118 ± 0.008	0.119 ± 0.010	$\mathbf{0.115 \pm 0.009}$
6	0.123 ± 0.007	0.123 ± 0.007	0.127 ± 0.008	0.123 ± 0.008	0.124 ± 0.010	0.127 ± 0.008	0.124 ± 0.010	0.123 ± 0.007	0.123 ± 0.007	0.123 ± 0.007	0.124 ± 0.009	$\mathbf{0.118 \pm 0.010}$
7	$\mathbf{0.126 \pm 0.008}$	$\mathbf{0.126 \pm 0.008}$	0.132 ± 0.009	0.128 ± 0.008	0.130 ± 0.010	0.133 ± 0.009	0.127 ± 0.009	$\mathbf{0.126 \pm 0.008}$	$\mathbf{0.126 \pm 0.008}$	$\mathbf{0.126 \pm 0.009}$	0.127 ± 0.009	$\mathbf{0.126 \pm 0.009}$
8	0.133 ± 0.010	0.133 ± 0.010	0.139 ± 0.011	0.132 ± 0.009	0.136 ± 0.010	0.137 ± 0.010	0.133 ± 0.010	0.133 ± 0.010	0.133 ± 0.010	0.133 ± 0.010	0.133 ± 0.010	$\mathbf{0.131 \pm 0.010}$
9	$\mathbf{0.138 \pm 0.012}$	$\mathbf{0.138 \pm 0.012}$	0.141 ± 0.012	0.139 ± 0.012	0.141 ± 0.013	0.142 ± 0.011	$\mathbf{0.138 \pm 0.012}$	$\mathbf{0.138 \pm 0.012}$	$\mathbf{0.138 \pm 0.012}$	$\mathbf{0.138 \pm 0.012}$	$\mathbf{0.138 \pm 0.012}$	$\mathbf{0.138 \pm 0.012}$
10	$\mathbf{0.142 \pm 0.013}$	$\mathbf{0.142 \pm 0.013}$	0.145 ± 0.014	0.145 ± 0.014	0.145 ± 0.014	0.147 ± 0.013	$\mathbf{0.142 \pm 0.013}$	$\mathbf{0.142 \pm 0.013}$	$\mathbf{0.142 \pm 0.013}$	$\mathbf{0.142 \pm 0.013}$	$\mathbf{0.142 \pm 0.013}$	0.144 ± 0.014
11	$\mathbf{0.151 \pm 0.016}$	$\mathbf{0.151 \pm 0.016}$	$\mathbf{0.151 \pm 0.016}$	$\mathbf{0.151 \pm 0.016}$	$\mathbf{0.151 \pm 0.016}$	$\mathbf{0.151 \pm 0.016}$	$\mathbf{0.151 \pm 0.016}$	$\mathbf{0.151 \pm 0.016}$	$\mathbf{0.151 \pm 0.016}$	$\mathbf{0.151 \pm 0.016}$	$\mathbf{0.151 \pm 0.016}$	$\mathbf{0.151 \pm 0.016}$

Table 17: Test error probability on AG News with majority voting (MV) aggregation for baselines versus MAP aggregation for Greedy MI. Results averaged over 30 evaluations. Bold indicates the lowest displayed mean error for each k, including ties.

k	Top-k (MV)	Term 1 (MV)	mRMR (MV)	TMC-Shapley (MV)	NaiveBayes Fwd (MV)	LLM-TOPLA (MV)	Society Entropy (MV)	Disagreement (MV)	KW Variance (MV)	Kuncheva (MV)	PCDM (MV)	Greedy MI (MAP)
1	$\mathbf{0.118 \pm 0.009}$	$\mathbf{0.118 \pm 0.009}$	$\mathbf{0.118 \pm 0.009}$	0.152 ± 0.026	$\mathbf{0.118 \pm 0.009}$	$\mathbf{0.118 \pm 0.009}$	$\mathbf{0.118 \pm 0.009}$	$\mathbf{0.118 \pm 0.009}$	$\mathbf{0.118 \pm 0.009}$	$\mathbf{0.118 \pm 0.009}$	$\mathbf{0.118 \pm 0.009}$	$\mathbf{0.118 \pm 0.009}$
2	0.147 ± 0.011	0.153 ± 0.007	0.241 ± 0.021	0.168 ± 0.015	0.153 ± 0.007	0.265 ± 0.009	0.148 ± 0.012	0.147 ± 0.011	0.147 ± 0.011	0.153 ± 0.007	0.153 ± 0.007	$\mathbf{0.112 \pm 0.010}$
3	0.137 ± 0.010	0.132 ± 0.011	0.156 ± 0.007	0.146 ± 0.008	0.148 ± 0.009	0.158 ± 0.007	0.132 ± 0.011	0.132 ± 0.011	0.132 ± 0.011	0.132 ± 0.011	0.132 ± 0.011	$\mathbf{0.111 \pm 0.010}$
4	0.146 ± 0.011	0.149 ± 0.007	0.159 ± 0.008	0.158 ± 0.008	0.159 ± 0.013	0.160 ± 0.010	0.146 ± 0.011	0.146 ± 0.011	0.146 ± 0.011	0.146 ± 0.011	0.146 ± 0.011	$\mathbf{0.112 \pm 0.009}$
5	0.143 ± 0.012	0.142 ± 0.011	0.161 ± 0.008	0.151 ± 0.007	0.149 ± 0.009	0.162 ± 0.007	0.142 ± 0.010	0.143 ± 0.012	0.143 ± 0.012	0.142 ± 0.012	0.142 ± 0.009	$\mathbf{0.115 \pm 0.009}$
6	0.155 ± 0.009	0.155 ± 0.009	0.172 ± 0.009	0.157 ± 0.007	0.156 ± 0.010	0.163 ± 0.008	0.153 ± 0.009	0.155 ± 0.009	0.155 ± 0.009	0.155 ± 0.010	0.153 ± 0.009	$\mathbf{0.118 \pm 0.010}$
7	0.152 ± 0.008	0.152 ± 0.008	0.157 ± 0.008	0.152 ± 0.008	0.155 ± 0.008	0.157 ± 0.008	0.150 ± 0.008	0.152 ± 0.008	0.152 ± 0.008	0.152 ± 0.008	0.151 ± 0.008	$\mathbf{0.126 \pm 0.009}$
8	0.159 ± 0.007	0.161 ± 0.007	0.166 ± 0.007	0.161 ± 0.008	0.167 ± 0.007	0.166 ± 0.007	0.159 ± 0.007	0.159 ± 0.007	0.159 ± 0.007	0.159 ± 0.007	0.159 ± 0.007	$\mathbf{0.131 \pm 0.010}$
9	0.158 ± 0.008	0.158 ± 0.008	0.163 ± 0.008	0.159 ± 0.008	0.163 ± 0.008	0.164 ± 0.008	0.158 ± 0.008	0.158 ± 0.008	0.158 ± 0.008	0.158 ± 0.008	0.158 ± 0.008	$\mathbf{0.138 \pm 0.012}$
10	0.165 ± 0.008	0.165 ± 0.008	0.169 ± 0.007	0.165 ± 0.009	0.167 ± 0.008	0.169 ± 0.007	0.165 ± 0.008	0.165 ± 0.008	0.165 ± 0.008	0.165 ± 0.008	0.165 ± 0.008	$\mathbf{0.144 \pm 0.014}$
11	0.165 ± 0.009	0.165 ± 0.009	0.165 ± 0.009	0.165 ± 0.009	0.165 ± 0.009	0.165 ± 0.009	0.165 ± 0.009	0.165 ± 0.009	0.165 ± 0.009	0.165 ± 0.009	0.165 ± 0.009	$\mathbf{0.151 \pm 0.016}$

Table 18: Test error probability on AG News with weighted majority voting (WV) aggregation for baselines versus MAP aggregation for Greedy MI. Results averaged over 30 evaluations. Bold indicates the lowest displayed mean error for each k, including ties.

k	Top-k (WV)	Term 1 (WV)	mRMR (WV)	TMC-Shapley (WV)	NaiveBayes Fwd (WV)	LLM-TOPLA (WV)	Society Entropy (WV)	Disagreement (WV)	KW Variance (WV)	Kuncheva (WV)	PCDM (WV)	Greedy MI (MAP)
1	$\mathbf{0.118 \pm 0.009}$	$\mathbf{0.118 \pm 0.009}$	$\mathbf{0.118 \pm 0.009}$	0.128 ± 0.020	$\mathbf{0.118 \pm 0.009}$	$\mathbf{0.118 \pm 0.009}$	$\mathbf{0.118 \pm 0.009}$	$\mathbf{0.118 \pm 0.009}$	$\mathbf{0.118 \pm 0.009}$	$\mathbf{0.118 \pm 0.009}$	$\mathbf{0.118 \pm 0.009}$	$\mathbf{0.118 \pm 0.009}$
2	0.122 ± 0.010	$\mathbf{0.112 \pm 0.010}$	0.125 ± 0.010	0.126 ± 0.006	0.113 ± 0.010	0.122 ± 0.010	$\mathbf{0.112 \pm 0.010}$	$\mathbf{0.112 \pm 0.011}$	0.122 ± 0.010	$\mathbf{0.112 \pm 0.010}$	$\mathbf{0.112 \pm 0.010}$	$\mathbf{0.112 \pm 0.010}$
3	0.123 ± 0.012	0.117 ± 0.010	0.115 ± 0.010	0.123 ± 0.007	0.116 ± 0.010	0.115 ± 0.010	0.117 ± 0.010	0.117 ± 0.010	0.117 ± 0.010	0.117 ± 0.010	0.117 ± 0.010	$\mathbf{0.111 \pm 0.010}$
4	0.127 ± 0.010	0.129 ± 0.008	0.134 ± 0.011	0.135 ± 0.010	0.128 ± 0.010	0.121 ± 0.008	0.127 ± 0.010	0.127 ± 0.010	0.127 ± 0.010	0.127 ± 0.010	0.127 ± 0.010	$\mathbf{0.112 \pm 0.009}$
5	0.129 ± 0.010	0.129 ± 0.010	0.137 ± 0.011	0.134 ± 0.008	0.130 ± 0.009	0.134 ± 0.010	0.130 ± 0.008	0.129 ± 0.010	0.129 ± 0.010	0.129 ± 0.010	0.130 ± 0.008	$\mathbf{0.115 \pm 0.009}$
6	0.128 ± 0.011	0.128 ± 0.011	0.136 ± 0.009	0.132 ± 0.008	0.128 ± 0.009	0.134 ± 0.008	0.129 ± 0.010	0.128 ± 0.011	0.128 ± 0.011	0.128 ± 0.011	0.128 ± 0.009	$\mathbf{0.118 \pm 0.010}$
7	0.129 ± 0.010	0.129 ± 0.010	0.132 ± 0.009	0.130 ± 0.012	0.130 ± 0.008	0.131 ± 0.009	0.129 ± 0.010	0.129 ± 0.010	0.129 ± 0.010	0.129 ± 0.010	0.129 ± 0.010	$\mathbf{0.126 \pm 0.009}$
8	0.135 ± 0.010	0.135 ± 0.009	$\mathbf{0.131 \pm 0.010}$	0.133 ± 0.010	$\mathbf{0.131 \pm 0.010}$	0.133 ± 0.010	0.135 ± 0.010	0.135 ± 0.010	0.135 ± 0.010	0.135 ± 0.010	0.135 ± 0.010	$\mathbf{0.131 \pm 0.010}$
9	0.137 ± 0.011	0.137 ± 0.011	0.138 ± 0.010	0.138 ± 0.012	0.137 ± 0.010	$\mathbf{0.135 \pm 0.011}$	0.137 ± 0.011	0.137 ± 0.011	0.137 ± 0.011	0.137 ± 0.011	0.137 ± 0.011	0.138 ± 0.012
10	$\mathbf{0.137 \pm 0.011}$	$\mathbf{0.137 \pm 0.011}$	0.140 ± 0.010	0.138 ± 0.011	0.138 ± 0.010	0.139 ± 0.009	$\mathbf{0.137 \pm 0.011}$	$\mathbf{0.137 \pm 0.011}$	$\mathbf{0.137 \pm 0.011}$	$\mathbf{0.137 \pm 0.011}$	$\mathbf{0.137 \pm 0.011}$	0.144 ± 0.014
11	$\mathbf{0.139 \pm 0.010}$	$\mathbf{0.139 \pm 0.010}$	$\mathbf{0.139 \pm 0.010}$	$\mathbf{0.139 \pm 0.010}$	$\mathbf{0.139 \pm 0.010}$	$\mathbf{0.139 \pm 0.010}$	$\mathbf{0.139 \pm 0.010}$	$\mathbf{0.139 \pm 0.010}$	$\mathbf{0.139 \pm 0.010}$	$\mathbf{0.139 \pm 0.010}$	$\mathbf{0.139 \pm 0.010}$	0.151 ± 0.016

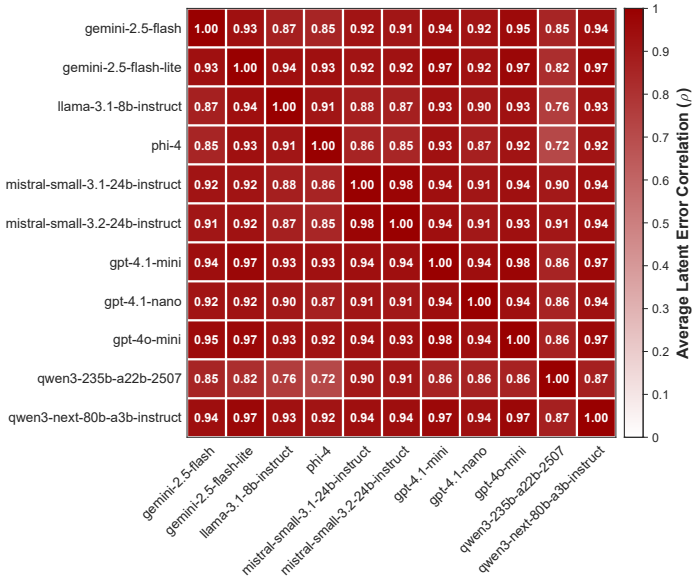


Figure 27: Averaged corr. matrix for AG News Classification dataset for 3 temperature settings.

Table 19: Model pool for AG News: 11 LLMs with individual accuracies and Gaussian-copula parameters. Results averaged over 3 temperature settings. Models sorted by accuracy.

Model	**Accuracy**	**Threshold** τ_j	**Avg.** ρ_j
qwen/qwen3-235b-a22b-2507	$87.9\% \pm 0.1\%$	-1.17	0.84
mistralai/mistral-small-3.2-24b-instruct	$84.9\% \pm 0.2\%$	-1.03	0.92
mistralai/mistral-small-3.1-24b-instruct	$84.1\% \pm 0.0\%$	-1.00	0.92
google/gemini-2.5-flash	$83.9\% \pm 0.1\%$	-0.99	0.91
openai/gpt-4o-mini	$82.3\% \pm 0.1\%$	-0.93	0.94
openai/gpt-4.1-mini	$82.3\% \pm 0.0\%$	-0.93	0.94
qwen/qwen3-next-80b-a3b-instruct	$82.1\% \pm 0.1\%$	-0.92	0.94
openai/gpt-4.1-nano	$81.4\% \pm 0.1\%$	-0.89	0.91
google/gemini-2.5-flash-lite	$79.8\% \pm 0.1\%$	-0.83	0.93
meta-llama/llama-3.1-8b-instruct	$76.9\% \pm 0.1\%$	-0.73	0.89
microsoft/phi-4	$72.9\% \pm 0.5\%$	-0.61	0.88
Average	81.7%	—	$\bar{\rho} = 0.91$

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Yes. The abstract and introduction explicitly outline our core claims: modeling LLM errors with a Gaussian-copula, establishing an error floor for Naive Bayes aggregation of correlated ensembles, and proposing the Greedy MI selection algorithm. These claims accurately reflect the paper's scope and are directly supported by the theoretical proofs in Section 3 and the extensive empirical evaluations across four datasets (MEDMCQA, MMLU, AG News, and IMDB) detailed in Section 5.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss it in a dedicated "Limitations and Discussion" subsection in Section 6.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: All assumptions are explicitly stated within the theorem statements in Section 3. We provide full, step-by-step mathematical proofs for all of our theoretical results in the supplementary material, specifically Appendices B,C,D, E,F.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: All information necessary to reproduce our results is thoroughly documented. Section 5 details the evaluated LLMs, prompt templates, and dataset splits. Furthermore, Appendix G provides comprehensive pseudo-code for our proposed Greedy MI algorithm, all baseline selection strategies, the aggregation methods, and the Gaussian-copula fitting procedure.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: To support verification of the reported results, we provide the source code and experimental datasets in the project repository.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Because our method involves ensemble selection rather than model training, traditional optimizers are not applicable. However, we fully specify all relevant evaluation parameters in Section 5 and in the Appendix, including API sampling temperatures (0.01, 0.3, 0.7), train/test split ratios, the number of independent runs, and the smoothing hyperparameters applied in the algorithms.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

Justification: We report the standard deviation across all our experiments to account for variability. In our plots, the shaded regions represent the standard deviation, and in our supplementary tables we explicitly report the mean $\pm$ standard deviation. These statistics are calculated over 30 independent evaluations per dataset, capturing the variance across different API sampling temperatures, independent runs, and randomized data splits.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We specify in the experimental setup and appendices that all LLM inference was conducted via the OpenRouter API. Furthermore, we detail that the subsequent data analysis, including mutual information estimation, copula fitting, and ensemble aggregation, requires minimal computational overhead and was executed locally on a standard personal computer.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: have reviewed the NeurIPS Code of Ethics and confirm that our research, methodology, and submission fully comply with its guidelines.

   Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We address the broader implications of our research in the Impact Statement section. We highlight the positive societal impacts, such as reducing the computational and environmental costs of LLM deployments, and clarify that our theoretical framework does not introduce new data collection mechanisms or direct ethical risks.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: All benchmark datasets utilized in our experiments (MEDMCQA, MMLU, AG News, and IMDB) are publicly available, with details provided in Section 5.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.

- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All public benchmark datasets (MEDMCQA, MMLU, AG News, and IMDB) and the foundation models evaluated in our ensembles are appropriately credited and cited within the Experiments section (Section 5) and the References.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: This work does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: We did not utilize crowdsourcing or conduct human subject research in this study.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: We did not utilize crowdsourcing or conduct human subject research in this study.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: Because evaluating LLM ensembles is the central scientific contribution of this work. We fully declare all LLMs queried during our experiments, alongside the evaluation protocols and exact prompt templates used across our benchmark datasets, in Section 5 and the corresponding appendices.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.