Merge Theory for LLMs: Set-Valued Inverses, Consensus Gains, and Uncertainty

Milad Shaddelan
milad.shaddelan@gmail.com

August 15, 2025

Abstract

We study the problem of merging predictive models into set-valued predictors when multiple inputs yield identical outputs. This non-injective structure breaks standard assumptions and causes systematic errors in abductive reasoning. We present Merge-Reasoner, a framework combining forward aggregation (MERGE) and backward enumeration (SPREAD). Our results characterize when thresholding, top-k selection, and consensus voting achieve Bayes-optimality, and when calibration is required for minimal coverage. We prove five main theorems: threshold-optimality under regularized risk, top-k Bayes-optimality under cardinality constraints, posterior bounds via merge degree under exchangeable priors, consensus gains with exponential error decay, and posterior optimality under proper scoring rules with minimal-cardinality coverage sets. The merge degree—preimage cardinality estimated via effective cluster count—provides principled uncertainty quantification. We separate theoretical guarantees from engineering heuristics, providing practical guidance for GPU optimization, diversity injection, and threshold calibration. Applications include medical diagnosis, multi-hop reasoning, and any domain where many-to-one mappings naturally arise.

1 Introduction

Specialized LLMs proliferate across domains, yet combining their capabilities remains theoretically ad hoc. Non-injective mappings—where multiple inputs yield identical outputs—break assumptions underlying current merging methods, causing systematic failures in abductive reasoning and diagnostic tasks. For example, in medical diagnosis, multiple diseases can present identical symptoms, making single-path reasoning insufficient.

This paper develops rigorous theoretical foundations for LLM merging through set-valued optimization. We formalize the problem mathematically (Section 3), derive practical algorithms (Section 5), and provide engineering guidelines (Section 6). We make four contributions:

- **We formalize** merging as finding set-valued inverses that capture the many-to-one structure of model predictions, proving optimality under regularized risk (Theorem 3) and cardinality constraints (Theorem 5).
- **We prove** that under exchangeable priors, merge degree bounds maximum posterior confidence (Theorem 8), and under conditional independence with accuracy $p > 1/2$, consensus voting achieves exponential error reduction (Theorem 11).
- **We establish** that posterior distributions minimize any strictly proper scoring rule, and we derive minimal-cardinality sets that achieve target coverage levels (Theorem 14 and Theorem 16).
- **We design** the Merge-Reasoner framework with Forward MERGE and Backward SPREAD algorithms, providing practical implementation guidelines for GPU optimization, diversity injection, and threshold calibration.

Our framework establishes the mathematical foundations for principled model merging, with clear theoretical guarantees and practical algorithmic guidelines.

Our contributions are formalized in five theorems. Theorem 3 establishes threshold-optimality under regularized risk. Theorem 5 gives top-k Bayes-optimality under cardinality constraints. Theorem 8 bounds posterior confidence via merge degree under exchangeable priors. Theorem 11 shows consensus gains with exponential error decay. Theorem 14 and Theorem 16 prove posterior optimality under proper scoring rules together with minimal-cardinality coverage sets.

2 Related Work

Our work builds upon several areas of research in machine learning and optimization theory.

2.1 Model Merging and Ensemble Methods

Traditional ensemble methods [Dietterich, 2000] combine predictions from multiple models through averaging or voting. Recent work on model merging [Wortsman et al., 2022] extends these ideas to parameter space, directly combining model weights. Task arithmetic [Ilharco et al., 2023] demonstrates that model capabilities can be manipulated through vector arithmetic in parameter space. Other approaches include Fisher-weighted averaging [Matena and Raffel, 2022], TIES-Merging [Yadav et al., 2024], and LoRA merging [Hu et al., 2022] for parameter-efficient fine-tuning.

2.2 Multi-Task Learning and Transfer Learning

Multi-task learning [Caruana, 1997] trains a single model on multiple tasks simultaneously. Our framework can be viewed as a post-hoc multi-task learning approach where pre-trained models are combined rather than jointly trained.

2.3 Set-Valued Analysis

Set-valued mappings have been extensively studied in optimization [Aubin and Frankowska, 2009] and control theory [Aubin, 1991]. We adapt these concepts to the model merging context, treating the merge operation as finding appropriate set-valued inverses. Related work on conformal prediction [Vovk et al., 2005, Romano et al., 2019] provides calibrated coverage guarantees for set predictions, which inspired our minimal-cardinality coverage sets (Theorem 16).

2.4 Uncertainty Quantification in Neural Networks

Bayesian neural networks [MacKay, 1992] provide a principled approach to uncertainty quantification. Deep ensembles [Lakshminarayanan et al., 2017] offer a practical alternative. Our uncertainty quantification methods combine insights from both approaches. The exponential error reduction in our consensus voting (Theorem 11) is supported by concentration inequalities [Hoeffding, 1963].

3 Methodology

3.1 Problem Formulation

Throughout, $M : \mathcal{X} \to \Delta(\mathcal{Y})$ denotes a model mapping inputs to distributions, and $\mathcal{D} = \{(x_j, y_j)\}_{j=1}^{n} \subseteq \mathcal{X} \times \mathcal{Y}$ is the dataset. Let $\mathcal{M} = \{M_1, M_2, \ldots, M_k\}$ be a collection of pre-trained LLMs, where each $M_i : \mathcal{X} \to \Delta(\mathcal{Y})$ maps inputs to predictive distributions over a finite output space $\mathcal{Y}$ (e.g., answer labels or normalized responses). We seek to find a merged model $M^* : \mathcal{X} \to 2^{\mathcal{Y}}$ that combines the capabilities of all models.

3.2 Set-Valued Inverses

Definition 1 (Set-Valued Merge Operator)**.** *The merge operator* $\Phi : \mathcal{M} \to (\mathcal{X} \to 2^{\mathcal{Y}})$ *is defined as:*

$$\Phi(\mathcal{M})(x) = \bigcup_{i=1}^{k} \{y \in \mathcal{Y} : P_{M_i}(y|x) > \tau_{min}\}$$

where τ_{min} *is a minimum probability threshold (e.g., 0.01) to define the truncated support. Here* $P_{M_i}(y|x)$ *denotes the conditional distribution emitted by* M_i *(e.g., normalized token/logit scores mapped into* $\mathcal{Y}$*).*

This formulation captures the multi-modal nature of merged predictions, acknowledging that different models may have valid but distinct outputs for the same input.

3.3 Consensus Gains

Definition 2 (Consensus Gain)**.** *The consensus gain* $G : \mathcal{M} \times \mathcal{X} \to \mathbb{R}^+$ *measures the agreement between models:*

$$G(\mathcal{M}, x) = \frac{1}{k(k-1)} \sum_{i \neq j} sim(M_i(x), M_j(x))$$

where $sim(\cdot, \cdot) \in [0, 1]$ *is a normalized similarity measure. In practice, we use cosine similarity of answer embeddings for text outputs or exact match indicator (1 if identical, 0 otherwise) for discrete labels.*

High consensus gain indicates that models agree on the output, suggesting reliable merged predictions.

3.4 Optimization Framework

We formulate model merging as the following optimization problem:

$$\theta^* = \arg\min_{\theta} \sum_{(x,y) \in \mathcal{D}} [\mathcal{L}(\Phi_\theta(\mathcal{M})(x), y) - \lambda G(\Phi_\theta(\mathcal{M}), x)]$$

where $\mathcal{L}$ is a loss function (typically cross-entropy or 0–1 loss surrogate) and $\lambda \in [0, 1]$ controls the importance of consensus.

3.5 Theoretical Foundations

We now present our main theoretical results that establish the optimality of threshold-based set predictors.

This first result formalizes the intuition that thresholding probabilities is risk-optimal under convex surrogates.

Theorem 3 (T1: Threshold Optimality via Regularization)**.** *Fix* $\lambda \in [0, 1]$*. Consider the regularized risk*

$$R_\lambda(S) = \sum_y P(y) \left(1 - \sum_{x \in S(y)} P(x|y)\right) + \lambda \sum_y P(y)|S(y)| \tag{1}$$

The set predictor that minimizes this regularized risk is:

$$S_\lambda(y) = \{x \in \mathcal{X} : P(x|y) \geq \lambda\} \tag{2}$$

Proof. **Step 1: Decompose the regularized risk.** We can rewrite the regularized risk as:

$$R_\lambda(S) = \sum_y P(y) \left[1 - \sum_{x \in S(y)} P(x|y) + \lambda |S(y)| \right] \tag{3}$$

Step 2: Pointwise optimization. Since the risk decomposes additively over y, we can minimize separately for each y:

$$\min_{S(y) \subseteq \mathcal{X}} \left[1 - \sum_{x \in S(y)} P(x|y) + \lambda |S(y)| \right] \tag{4}$$

Step 3: Rewrite per-element contribution. Note that $|S(y)| = \sum_{x \in S(y)} 1$, so:

$$1 - \sum_{x \in S(y)} P(x|y) + \lambda |S(y)| = 1 + \sum_{x \in S(y)} (\lambda - P(x|y)) \tag{5}$$

Step 4: Determine optimal inclusion. Since the constant 1 doesn't affect optimization, we minimize:

$$\sum_{x \in S(y)} (\lambda - P(x|y)) \tag{6}$$

Each element x contributes $(\lambda - P(x|y))$ to this sum. Therefore:

- Include x if $\lambda - P(x|y) < 0$, i.e., $P(x|y) > \lambda$
- Exclude x if $\lambda - P(x|y) > 0$, i.e., $P(x|y) < \lambda$
- Either choice is optimal if $P(x|y) = \lambda$

Step 5: Conclude. The optimal set is $S_\lambda(y) = \{x \in \mathcal{X} : P(x|y) \geq \lambda\}$, where ties at $P(x|y) = \lambda$ can be broken arbitrarily without affecting the risk. □

Corollary 4 (Comparison with Singleton Predictors)**.** *Any singleton predictor $\{\hat{x}(y)\}$ is equivalent to using some threshold $\tau \geq \max_{x \neq \hat{x}(y)} P(x|y)$. The best singleton predictor chooses $\hat{x}(y) \in \arg\max_x P(x|y)$ and has risk:*

$$R_{singleton} = 1 - \sum_y P(y) \max_x P(x|y) \tag{7}$$

For any $\lambda \leq \max_x P(x|y)$, the threshold set $S_\lambda(y)$ contains at least this optimal singleton, so $R_\lambda(S_\lambda) \leq R_{singleton}$ (comparing only the miss probability component).

The following theorem shows that selecting the top-k highest-probability elements minimizes Bayes risk under cardinality constraints.

Theorem 5 (T2: Top-k Bayes Optimality under 0–1 Miss Loss)**.** *Let $\mathcal{X}$ be finite and let $P(x|y)$ be a posterior on $\mathcal{X}$ for each observation y. For a fixed $k \in \mathbb{N}$, among all set predictors $S(y) \subseteq \mathcal{X}$ with $|S(y)| \leq k$, the Bayes risk*

$$R(S) = \sum_y P(y) \left(1 - \sum_{x \in S(y)} P(x|y) \right) \tag{8}$$

is minimized by any $S_k(y)$ containing the k largest posterior-probability elements (breaking ties arbitrarily).

Proof. The risk decomposes over y, so we minimize pointwise:

$$\max_{|S(y)|\leq k} \sum_{x\in S(y)} P(x|y) \tag{9}$$

If $x \in S$ and $x' \notin S$ with $P(x'|y) > P(x|y)$, replacing x by x' increases the objective. Iterating yields a set of the k largest posteriors (with arbitrary tie-breaking).

If fewer than k elements have positive posterior, any superset up to size k that adds only zero-probability elements is equally optimal. Summing over y preserves optimality. □

Corollary 6 (Singleton Optimality)**.** *For $k = 1$, the optimal predictor selects $\hat{x}(y) \in \arg\max_x P(x|y)$ with risk*

$$R = 1 - \sum_y P(y) \max_x P(x|y) \tag{10}$$

Remark 7 (Connection to Knapsack Problem)**.** *The top-k selection problem can be viewed as a 0–1 knapsack problem with:*

- *Capacity: k (maximum set size)*
- *Item weights: all equal to 1*
- *Item values: posterior probabilities $P(x|y)$*

The greedy solution (selecting highest-value items first) is optimal due to uniform weights, yielding the top-k posterior elements.

Note: Ties at the k-th position can be broken arbitrarily without affecting optimality.

This theorem characterizes how non-injective mappings create fundamental uncertainty bounds.

Theorem 8 (T3: Uncertainty via Merge Degree)**.** *Let $\mathcal{X}$ be finite and $f : \mathcal{X} \to \mathcal{Y}$ be a deterministic mapping (Assumptions 1-2). For $y \in \mathcal{Y}$, define the merge degree $m_f(y) = |f^{-1}(y)|$. Under Assumption 3 (exchangeable priors within each preimage set $f^{-1}(y)$ with $P(y) > 0$), the maximum posterior probability*

$$\max_x P(x|y) = \frac{1}{m_f(y)} \tag{11}$$

is strictly decreasing in the merge degree $m_f(y)$.

Proof. Since f is deterministic, $P(y|x) = \mathbf{1}\{f(x) = y\}$.

By exchangeability within $f^{-1}(y)$, there exists $\pi_y > 0$ such that $P(x) = \pi_y$ for all $x \in f^{-1}(y)$.

The marginal probability is

$$P(y) = \sum_{x\in f^{-1}(y)} P(x) = m_f(y) \cdot \pi_y \tag{12}$$

For $x \in f^{-1}(y)$, the posterior is:

$$P(x|y) = \frac{P(y|x)P(x)}{P(y)} = \frac{1 \cdot \pi_y}{m_f(y) \cdot \pi_y} = \frac{1}{m_f(y)} \tag{13}$$

Thus the posterior is uniform over $f^{-1}(y)$, yielding:

$$\max_x P(x|y) = \frac{1}{m_f(y)} \tag{14}$$

This is strictly decreasing when $m_f(y)$ increases by adding elements with positive prior (constant only if adding zero-prior elements).

The uniform distribution is majorized by any other distribution on the same support, reinforcing that spreading mass over more elements reduces the maximum. □

Remark 9 (Interpretation)**.** *In a lossy many-to-one world, the more distinct inputs can yield the same observation y (higher merge degree), the smaller any single input's posterior chance must be; with an exchangeable prior, it's exactly $1/m_f(y)$.*

Remark 10 (General Case without Exchangeability)**.** *Without exchangeability, if $A = f^{-1}(y)$, $S = \sum_{x\in A} P(x)$, and $p_{\max} = \max_{x\in A} P(x)$, then $\max_x P(x|y) = p_{\max}/S$. Adding a new element with prior $q \le p_{\max}$ gives:*

$$\frac{\max(p_{\max}, q)}{S+q} \le \frac{p_{\max}}{S} \tag{15}$$

with strict decrease when $0 < q < p_{\max}$. Note that if $q > p_{\max}$, the maximum posterior can increase. The exchangeability assumption ensures all elements contribute equally, guaranteeing monotone decrease.

This theorem quantifies the classical wisdom that majority vote is better than any single classifier, except in the degenerate $n = 2$ case.

Theorem 11 (T4: Forward Consensus Gain)**.** *Under Assumption 4 (conditional independence of reasoning chains), for $n \ge 3$ chains each producing a correct conclusion with probability $p > \frac{1}{2}$, and incorrect otherwise. If all chains map to the same binary decision space, the majority vote output is correct with probability strictly greater than p (for $n = 2$ with fair tie-breaking, it equals p). Moreover, as $n \to \infty$, the probability tends to 1.*

Proof. **Step 1: Binomial Distribution Setup.** Let X_i be the indicator variable for the i-th reasoning chain:

- $X_i = 1$ if chain i produces the correct conclusion
- $X_i = 0$ if chain i produces the incorrect conclusion

Since each chain has probability p of being correct and chains are independent:

$$P(X_i = 1) = p \tag{16}$$
$$P(X_i = 0) = 1 - p \tag{17}$$

Let $S_n = \sum_{i=1}^{n} X_i$ be the total number of correct chains. By independence, S_n follows a binomial distribution:

$$S_n \sim \text{Binomial}(n, p) \tag{18}$$
$$P(S_n = k) = \binom{n}{k} p^k (1-p)^{n-k} \tag{19}$$

Step 2: Probability of Majority Correctness. The majority vote is correct when more than half the chains are correct. For odd n, this means at least $\frac{n+1}{2}$ correct chains. For even n, we break ties uniformly at random.

Define $M_n(p) = P(\text{majority correct with } n \text{ chains})$.

Step 3: Showing $P_{\text{majority}} > p$ for $p > \frac{1}{2}$. Let $q = 1-p$, where $p > \frac{1}{2}$. For odd $n = 2m+1$:

$$M_{2m+1}(p) = P(S_{2m+1} \ge m+1) \tag{20}$$

Key Identity (Monotonicity in n): For odd $n = 2m+1$, we have:

$$M_{2m+3}(p) - M_{2m+1}(p) = \binom{2m+2}{m+1} p^{m+1} q^{m+1} (2p-1) > 0 \tag{21}$$

This follows by expanding the binomial tails and using Pascal's rule. Since $p > \frac{1}{2}$, we have $2p - 1 > 0$, establishing monotonicity.

Base Case ($n = 3$): For $n = 3$:

$$M_3(p) = p^3 + 3p^2q = p^3 + 3p^2(1-p) = 3p^2 - 2p^3 \tag{22}$$

We need to show $M_3(p) > p$ for $p \in (\frac{1}{2}, 1)$.

$$\begin{aligned} M_3(p) - p &= 3p^2 - 2p^3 - p && (23)\\ &= p(-2p^2 + 3p - 1) && (24)\\ &= p(1-2p)(p-1) && (25) \end{aligned}$$

For $p \in (\frac{1}{2}, 1)$:

- $p > 0$
- $1 - 2p < 0$ (since $p > \frac{1}{2}$)
- $p - 1 < 0$ (since $p < 1$)
- Product of two negative terms is positive

Therefore $M_3(p) - p > 0$, so $M_3(p) > p$ for all $p \in (\frac{1}{2}, 1)$.
By monotonicity, $M_{2m+1}(p) \geq M_3(p) > p$ for all odd $n \geq 3$.

Step 4: Asymptotic Behavior using Central Limit Theorem. By the Central Limit Theorem, as $n \to \infty$:

$$\frac{S_n - np}{\sqrt{np(1-p)}} \xrightarrow{d} \mathcal{N}(0,1) \tag{26}$$

The majority is correct when $S_n > \frac{n}{2}$, which is equivalent to:

$$\frac{S_n - np}{\sqrt{np(1-p)}} > \frac{n/2 - np}{\sqrt{np(1-p)}} = -\sqrt{n}\frac{p - 1/2}{\sqrt{p(1-p)}} \tag{27}$$

Since $p > \frac{1}{2}$, we have $p - \frac{1}{2} > 0$. Therefore:

$$-\sqrt{n}\frac{p - 1/2}{\sqrt{p(1-p)}} \to -\infty \text{ as } n \to \infty \tag{28}$$

Thus:

$$P_{\text{majority}} = P\left(Z > -\sqrt{n}\frac{p - 1/2}{\sqrt{p(1-p)}}\right) \to P(Z > -\infty) = 1 \tag{29}$$

where $Z \sim \mathcal{N}(0,1)$.

Step 5: Exponential Error Decay via Chernoff Bound. For finite n, the Chernoff bound [Hoeffding, 1963] gives:

$$P(S_n \leq n/2) \leq \exp\left(-2n(p - 1/2)^2\right) \tag{30}$$

Thus the error probability decays exponentially in n with rate $2(p - 1/2)^2$. □

Corollary 12 (Condorcet's Jury Theorem)**.** *For a jury of n members, each with independent probability $p > \frac{1}{2}$ of making the correct decision, the probability of a correct majority decision increases monotonically with n and approaches certainty as $n \to \infty$.*

3.6 Proper Scoring Rules (Probabilistic) and Minimal-Cardinality Coverage Sets (Set-Valued)

Posterior predictions are not only intuitive, but in fact uniquely minimize every proper scoring rule.

We now establish the optimality of posterior-based predictions under proper scoring rules.

Definition 13 (Proper Scoring Rule)**.** *A scoring rule $S : \Delta(\mathcal{X}) \times \mathcal{X} \to \mathbb{R}$ assigns a score $S(q, x)$ when prediction q is made and outcome x occurs. The rule is* proper *if for any true distribution $p \in \Delta(\mathcal{X})$:*

$$\mathbb{E}_{x\sim p}[S(p, x)] \leq \mathbb{E}_{x\sim p}[S(q, x)] \quad \forall q \in \Delta(\mathcal{X}) \tag{31}$$

The rule is strictly proper *if equality holds only when $q = p$.*

Here $\Delta(\mathcal{X})$ denotes the probability simplex over $\mathcal{X}$ (the set of all probability distributions on the finite set $\mathcal{X}$).

Theorem 14 (T5a: Posterior-Optimality for Proper Scoring Rules)**.** *Let S be a strictly proper scoring rule. Given observation y, the posterior distribution $q^*(x|y) = P(x|y)$ uniquely minimizes the expected score:*

$$q^*(\cdot|y) = \arg\min_{q\in\Delta(\mathcal{X})} \mathbb{E}_{x\sim P(\cdot|y)}[S(q, x)] \tag{32}$$

Proof. Let $p(\cdot) = P(\cdot|y)$ denote the true posterior distribution given y. We need to show that for any $q \in \Delta(\mathcal{X})$:

$$\mathbb{E}_{x\sim p}[S(p, x)] \leq \mathbb{E}_{x\sim p}[S(q, x)] \tag{33}$$

with equality if and only if $q = p$.

This follows directly from the definition of strict propriety. Since S is strictly proper, for any true distribution p:

$$\mathbb{E}_{x\sim p}[S(p, x)] < \mathbb{E}_{x\sim p}[S(q, x)] \quad \text{for all } q \neq p \tag{34}$$

Applying this with $p = P(\cdot|y)$, we have:

$$\mathbb{E}_{x\sim P(\cdot|y)}[S(P(\cdot|y), x)] < \mathbb{E}_{x\sim P(\cdot|y)}[S(q, x)] \tag{35}$$

for all $q \neq P(\cdot|y)$.

Therefore, $q^*(\cdot|y) = P(\cdot|y)$ uniquely minimizes the expected score. □

Definition 15 (Calibrated Set)**.** *For coverage level $\alpha \in (0, 1)$, a prediction set $C(y) \subseteq \mathcal{X}$ is* calibrated *if:*

$$P(X \in C(y)|Y = y) \geq \alpha \tag{36}$$

Theorem 16 (T5b: Minimal-Cardinality Calibrated Sets)**.** *Fix y and $\alpha \in (0, 1)$. Among all sets $C(y) \subseteq \mathcal{X}$ satisfying the coverage constraint $\sum_{x\in C(y)} P(x|y) \geq \alpha$, the set that collects the largest posterior probabilities in descending order until the running sum first exceeds α has minimal cardinality.*

Proof. Sort the elements of $\mathcal{X}$ by their posterior probabilities:

$$P(x_1|y) \geq P(x_2|y) \geq \cdots \geq P(x_{|\mathcal{X}|}|y) \tag{37}$$

Let $k = \min\{j : \sum_{i=1}^{j} P(x_i|y) \geq \alpha\}$.

Claim: $C^*(y) = \{x_1, \ldots, x_k\}$ has minimal cardinality among all feasible sets.

For any feasible set C with $|C| < k$, its total probability mass is at most:

$$\sum_{x\in C} P(x|y) < \sum_{i=1}^{k-1} P(x_i|y) < \alpha \tag{38}$$

The first inequality follows because any set of size $< k$ cannot contain all of $\{x_1, \ldots, x_{k-1}\}$, and replacing any element outside this set with one inside can only increase the total mass. The second inequality follows from the definition of k. Note that the greedy choice is optimal due to the monotone submodular structure of the coverage function [Nemhauser et al., 1978].

Therefore, no set of size $< k$ can achieve coverage α, making $C^*(y)$ minimal. In the presence of ties at the cutoff index k, any set of minimal size that attains cumulative posterior mass $\geq \alpha$ is optimal. □

Remark 17 (Connection to Regularized Loss)**.** *For the regularized objective:*

$$R_\lambda(C) = 1 - \sum_{x \in C(y)} P(x|y) + \lambda |C(y)| \tag{39}$$

The optimal set is the threshold set $C_\lambda(y) = \{x : P(x|y) \geq \lambda\}$, connecting back to Theorem 3.

Remark 18 (Decision-Theoretic Interpretation)**.** *The optimality of posterior predictions reflects fundamental principles in decision theory:*

- ***For probabilistic predictions:*** *Honesty is optimal under proper scoring rules. The posterior $P(\cdot|y)$ represents our true beliefs given evidence y, and any deviation increases expected loss.*
- ***For set predictions:*** *When we want coverage at minimal size, we should include the most probable outcomes first. This greedily achieves the coverage constraint with the fewest elements.*

Both results show that Bayesian updating provides not just coherent but optimal predictions under natural loss functions and constraints.

4 The Merge-Reasoner Framework

4.1 Problem Setup

Consider a deterministic mapping $f : \mathcal{X} \to \mathcal{Y}$ with finite $\mathcal{X}$ and $P(y) > 0$ for all observable $y \in \mathcal{Y}$. The preimage $f^{-1}(y) = \{x \in \mathcal{X} : f(x) = y\}$ may contain multiple elements when f is non-injective. Define the **merge degree** $m_f(y) = |f^{-1}(y)|$ as the preimage cardinality. Conditional on observing y, and under exchangeable priors over $f^{-1}(y)$, the expected 0–1 miss loss equals $(m_f(y) - 1)/m_f(y)$, increasing with merge degree.

The posterior distribution given observation y is $P(x|y) = P(x) \cdot \mathbf{1}\{f(x) = y\}/P(y)$. Under exchangeable priors within $f^{-1}(y)$, we have $P(x|y) = 1/m_f(y)$ uniformly.

Mathematical Assumptions
Throughout this work, we assume:

1. **Finite domain:** $\mathcal{X}$ is finite and $P(y) > 0$ for all $y \in \mathcal{Y}$.
2. **Deterministic mapping:** $f : \mathcal{X} \to \mathcal{Y}$ is deterministic; the preimage $f^{-1}(y) = \{x \in \mathcal{X} : f(x) = y\}$ may be non-injective.
3. **Exchangeable priors for T3:** For all $x, x' \in f^{-1}(y)$, $P(x) = P(x')$, implying $P(x|y) = 1/m_f(y)$.
4. **Independence for T4:** Reasoning chains (or "voters") are conditionally independent given the true x, with per-chain accuracy $p > 1/2$.
5. **Merge degree estimation:** In practice, $m_f(y)$ is estimated via semantic clustering over generated chains; this is an approximation of the true preimage size.
6. **Approximation for large $\mathcal{X}$:** When $|\mathcal{X}|$ is large, top-k or thresholded enumeration approximates $f^{-1}(y)$ rather than fully enumerating it.

These assumptions are invoked explicitly where needed for the theoretical guarantees (T1–T5).

4.2 Framework Overview

Merge-Reasoner operates in two phases: **Forward MERGE** aggregates k reasoning chains into semantic clusters and selects via consensus scoring; **Backward SPREAD** enumerates high-posterior explanations for observation y and returns calibrated set $S(y)$. Merge degree drives uncertainty quantification throughout both phases.

5 Algorithms

5.1 Forward MERGE Algorithm

Algorithm 1 Forward MERGE

Require: Prompt p, Model $\mathcal{M}$, Chains $k = 8$, Temperature $T = 0.7$
Ensure: Answer, merge degree, vote margin, clusters

$chains \leftarrow \mathcal{M}$.GenerateBatch$(p, k, T)$ $\{O(k)$ with parallelization$\}$
$clusters \leftarrow$ SemanticCluster$(chains, \tau = 0.95)$ {Cosine similarity threshold, single-linkage clustering}
for each cluster c in clusters **do**
  $c.score \leftarrow$ LogSumExp([$ch.logprob/ch.length$ for ch in $c.members$]) {Length-normalized}
  $c.answer \leftarrow$ MajorityVote$(c.members)$
end for
$p_i \leftarrow |c_i|/k$ for each cluster c_i
$N_{\text{eff}} \leftarrow 1/\sum_i p_i^2$ {Effective number of distinct explanations}
$merge\_degree \leftarrow N_{\text{eff}}$ {Higher $N_{\text{eff}} \Rightarrow$ higher uncertainty}
$best\_cluster \leftarrow \arg\max_c c.score$
$vote\_margin \leftarrow |best\_cluster|/k$
return $\{answer : best\_cluster.answer, merge\_degree, vote\_margin, clusters\}$

Complexity: $O(k)$ generation + $O(k^2)$ clustering. **Note**: GPU speedup and KV cache claims are empirical observations requiring validation.

5.2 Backward SPREAD Algorithm

Algorithm 2 Backward SPREAD

Require: Observation y, Model $\mathcal{M}$, Either k or score threshold τ_{score}
Ensure: Explanations set, abstention flag

explanations $\leftarrow$ []
$T \leftarrow [0.3, 0.5, 0.7, 0.9]$ {Temperature schedule}
for each τ in T **do**
  batch $\leftarrow \mathcal{M}$.ExplainBatch$(y, k/4, \tau)$ {$O(k)$ generation}
  explanations.extend(*batch*)
end for
verified $\leftarrow$ []
for each exp in explanations **do**
  if Verifier.Check(exp, y) **then**
    verified.append$((exp, \mathcal{M}.\text{Score}(exp)))$
  end if
end for
if k is not None **then**
  selected $\leftarrow$ TopK$(verified, k)$ {Select k highest scores, descending}
else
  selected $\leftarrow \{(e, s) : (e, s) \in verified, s > \tau_{\text{score}}\}$ {Score threshold}
end if
final_set $\leftarrow$ SemanticDeduplicate(*selected*)
abstain $\leftarrow (|final\_set| = 0) \vee (\max_s s < \tau_{\text{conf}})$ {Abstain if empty or low confidence}
return $\{explanations : final\_set, abstain : abstain\}$

Complexity: $O(k)$ generation + $O(k)$ verification. **Note**: Batch verifier optimization is an empirical observation.

5.3 Theoretical Justification

Algorithmic components implement our theoretical guarantees:

- Top-k selection implements Theorem 5 (T2) for cardinality-constrained optimization
- Threshold selection implements Theorem 3 (T1); note that algorithmic τ_{score} approximates theoretical regularization parameter λ
- Merge degree estimation via N_{eff} approximates Theorem 8 (T3)'s bound $\max_x P(x|y) = 1/m_f(y)$
- Majority voting implements Theorem 11 (T4)'s consensus gain with exponential error reduction
- Posterior predictions implement Theorem 14 (T5a) for proper scoring rule optimality
- Calibrated sets implement Theorem 16 (T5b) for minimal-cardinality coverage at level α

6 Practical Engineering

6.1 GPU Optimization

Batch **k=4–8** chains for efficient parallelization. **Empirical observation (implementation detail):** Modern GPUs can achieve near-linear speedup up to $k = 8$ with proper KV cache

management; cache reuse across prompt variants may reduce per-chain latency.[1]

6.2 Early Exit Strategy

Stop generation when consensus margin exceeds threshold (75% agreement). Reduces computational cost while maintaining accuracy. Only applied if abstention predicate would remain false.

6.3 Diversity Injection

- Temperature variation: $\mathbf{T} \in [0.3, 0.9]$
- Random seed permutation across chains
- Prompt phrasing variations
- Avoid beam search (collapses to high-probability modes, reducing cluster variety)

6.4 Verifier Suite

Deploy heterogeneous verifiers:

- NLI: DeBERTa-v3 [He et al., 2023] for consistency checking
- Code: Unit test execution
- Math: SymPy symbolic verification
- Facts: Retrieval-based validation

6.5 Threshold Calibration

Plot risk-coverage curves on held-out validation data. Select λ for target empirical coverage α:

$$\lambda^* = \arg\min_{\lambda} \{R_\lambda(S_\lambda) : \text{EmpiricalCoverage}(S_\lambda) \geq \alpha\} \tag{40}$$

Note: This uses empirical validation coverage (implementation detail), not the oracle posterior from Theorem 16.

6.6 Output Contract

Output Structure:

- `'answer'`: `str` – Primary prediction
- `'merge_degree'`: `float` – N_{eff} (a.k.a. effective number of explanations)
- `'vote_margin'`: `float` – Consensus strength $[0, 1]$
- `'alt_explanations'`: `List[str]` – $S(y)$ alternatives
- `'cluster_id'`: `str` – Semantic cluster identifier
- `'abstain_flag'`: `bool` – Abstention decision

Abstention Rule: abstain = (vote_margin $< \tau_{\text{margin}}$) $\vee$ ($\max_c c.\text{score} < \tau_{\text{conf}}$)

[1] These heuristics are implementation-specific and not required by the theoretical results. Theoretical guarantees (T1–T5) hold under their stated assumptions, independent of implementation choices.

6.7 Reproducibility Specifications

- Seeds: Fixed per chain
- Context: 4096 tokens
- Defaults: $\mathbf{k=8}$, $\lambda=\mathbf{0.3}$ (theoretical), $\tau_{\mathbf{score}}=\mathbf{0.5}$ (algorithmic)
- Software: PyTorch 2.0, Transformers 4.30, NumPy 1.24
- Random seeds: Fixed per chain (seeds 0 to k-1)
- Hardware: A100 80GB
- Token budget: 10K/problem
- Optimizations: INT8, Flash Attention 2

7 Application Domains

7.1 Effective Application Domains

Table 1 summarizes domains where Merge-Reasoner provides significant improvements.

Table 1: Domains where Merge-Reasoner is effective

Domain	**Rationale**
Abductive reasoning	Many causes $\rightarrow$ one effect
Medical/fault diagnosis	Symptom aliasing
Multi-hop QA	Path convergence
Mathematical problems	Alternative solutions
Information retrieval	Near-duplicate merging
RAG with synonyms	Phrasing variants
Noisy measurements	Sensor aliasing
Lossy compression	Non-injective encoding

7.2 Ineffective Application Domains

Table 2 identifies scenarios where Merge-Reasoner provides limited benefit.

Table 2: Domains where Merge-Reasoner is ineffective

Domain	**Limitation**
Injective tasks	No merging possible
Poor base accuracy ($p \leq 0.5$)	Error amplification
Highly correlated chains	Independence violated
Miscalibrated k or λ	Poor selection
Infinite $\mathcal{X}$	Intractable enumeration

7.3 Safety Considerations and Failure Modes

Risks:

- Correlated errors from model biases
- Adversarial inputs gaming verifier agreement
- Semantic clustering failures on OOD text
- Semantic cluster poisoning via adversarial phrasing

Mitigations:

- Enforce diversity through explicit prompt variation
- Require agreement across heterogeneous verifiers
- Audit tool outputs for manipulation
- Set conservative abstention thresholds

7.4 Key Design Principles

What is right:

- Model non-injective structure explicitly
- Use merge degree for uncertainty quantification
- Leverage GPU batching for multi-chain inference
- Combine forward consensus with backward enumeration
- Calibrate thresholds on held-out data

What to avoid:

- Avoid applying to injective problems
- Avoid beam search (kills diversity)
- Avoid ignoring verifier disagreement
- Avoid arbitrary k or λ selection
- Avoid assuming independence with correlated prompts

8 Discussion

8.1 Theoretical Insights

Three key insights emerge from our theoretical framework:

1. **Non-convexity Intuition:** Set-valued formulation is motivated by the potential non-convexity of merge spaces where traditional weight averaging may fail. The set-valued approach can capture multiple distinct solutions that may lie in disconnected regions of parameter space, though proving this non-convexity formally remains future work.

2. **Consensus as Regularization:** Consensus gains (T4) act as implicit regularization. Under the independence assumption with per-chain accuracy $p > 1/2$, error reduction is exponential in the number of agreeing chains, with rate $2(p - 1/2)^2$ by the Chernoff bound (Theorem 11).

3. **Uncertainty via Merge Degree:** Under exchangeable priors, T3 establishes that uncertainty is inversely proportional to merge degree. Specifically, the maximum posterior probability equals exactly $1/|f^{-1}(y)|$, providing a principled calibration mechanism. Without exchangeability, the equality need not hold; in general $\max_x P(x|y) \geq 1/|f^{-1}(y)|$ and $\max_x P(x|y) \geq 1/N_{\text{eff}}$, with equality only in the uniform case.

8.2 Practical Implications

Our findings have several practical implications:

- **Model Selection:** Not all models benefit from merging. Our consensus gain metric can identify compatible models before merging.
- **Adaptive Merging:** The framework supports input-dependent merging weights, allowing the merged model to dynamically select which expertise to use.
- **Continual Learning:** The set-valued formulation naturally extends to continual learning scenarios where new models are incrementally merged.

8.3 Limitations and Future Work

Our guarantees require assumptions such as exchangeability (Theorem 8) and independence of errors across chains (Theorem 11). When these assumptions are violated—for instance, when reasoning chains share systematic biases or when priors are non-uniform—the theoretical bounds may not hold. Correlated errors remain an open challenge for future work.

Additional limitations include:

- **Empirical Validation:** While our theoretical guarantees are rigorously proven, comprehensive empirical validation across diverse benchmarks and model architectures remains future work.
- **Computational Scalability:** The computational cost scales with the number of models and chains, requiring optimization for practical deployment.
- **Similarity Measures:** The choice of similarity measure for clustering remains problem-dependent and requires further theoretical analysis.
- **Large Model Extension:** Extending the framework to very large models (>100B parameters) needs both theoretical and practical refinements.

Future work will focus on empirical validation of the theoretical predictions, developing efficient approximation algorithms, and extending the framework to multimodal and continual learning settings.

9 Conclusion

We formalized LLM merging through set-valued optimization, proving five theoretical guarantees that explain when and why merging succeeds. The framework's merge degree provides principled uncertainty quantification, while consensus voting enables exponential error reduction under

independence assumptions. Our theoretical analysis establishes conditions for optimal set selection, characterizes uncertainty bounds, and proves convergence properties.

Beyond immediate applications, our work impacts federated learning and continual model adaptation, where non-injective mappings naturally arise. As specialized LLMs proliferate, principled merging becomes critical infrastructure—our theoretical foundations enable reliable combination of diverse model capabilities.

We hope this framework helps bridge formal guarantees with practical ensemble engineering, and opens new directions in calibration, abstention, and consensus theory.

Code: `https://github.com/miladshd/merge-theory`

References

Jean-Pierre Aubin. *Viability theory*. Birkhäuser, 1991.

Jean-Pierre Aubin and Hélène Frankowska. *Set-valued analysis*. Springer Science & Business Media, 2009.

Rich Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997.

Thomas G Dietterich. Ensemble methods in machine learning. *International workshop on multiple classifier systems*, pages 1–15, 2000.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. Deberta-v3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *International Conference on Learning Representations*, 2023.

Wassily Hoeffding. *Probability inequalities for sums of bounded random variables*, volume 58. Taylor & Francis, 1963.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *International Conference on Learning Representations*, 2022.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. *International Conference on Learning Representations*, 2023.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

David JC MacKay. A practical bayesian framework for backpropagation networks. *Neural computation*, 4(3):448–472, 1992.

Michael S Matena and Colin A Raffel. Merging models with fisher-weighted averaging. *Advances in Neural Information Processing Systems*, 35:17703–17716, 2022.

George L Nemhauser, Laurence A Wolsey, and Marshall L Fisher. An analysis of approximations for maximizing submodular set functions—i. *Mathematical programming*, 14(1):265–294, 1978.

Yaniv Romano, Evan Patterson, and Emmanuel Candès. Conformalized quantile regression. *Advances in neural information processing systems*, 32, 2019.

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. Algorithmic learning in a random world. 2005.

Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, et al. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. *International Conference on Machine Learning*, pages 23965–23998, 2022.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. Ties-merging: Resolving interference when merging models. *Advances in Neural Information Processing Systems*, 36, 2024.

A Detailed Proofs

A.1 Proof of Theorem 3

Step 1: Decompose the regularized risk. We rewrite the regularized risk as:

$$R_\lambda(S) = \sum_y P(y) \left[1 - \sum_{x \in S(y)} P(x|y) + \lambda |S(y)| \right] \tag{41}$$

Step 2: Pointwise optimization. Since the risk decomposes additively over y, we minimize separately for each y:

$$\min_{S(y) \subseteq \mathcal{X}} \left[1 - \sum_{x \in S(y)} P(x|y) + \lambda |S(y)| \right] \tag{42}$$

Step 3: Rewrite per-element contribution. Note that $|S(y)| = \sum_{x \in S(y)} 1$, so:

$$1 - \sum_{x \in S(y)} P(x|y) + \lambda |S(y)| = 1 + \sum_{x \in S(y)} (\lambda - P(x|y)) \tag{43}$$

Step 4: Determine optimal inclusion. Since the constant 1 doesn't affect optimization, we minimize $\sum_{x \in S(y)} (\lambda - P(x|y))$. Each element x contributes $(\lambda - P(x|y))$ to this sum. Therefore:

- Include x if $P(x|y) > \lambda$
- Exclude x if $P(x|y) < \lambda$
- Either choice is optimal if $P(x|y) = \lambda$

Step 5: Conclude. The optimal set is $S_\lambda(y) = \{x \in \mathcal{X} : P(x|y) \geq \lambda\}$. □

A.2 Proof of Theorem 5

The risk decomposes over y: $\max_{|S(y)| \leq k} \sum_{x \in S(y)} P(x|y)$.

If $x \in S$ and $x' \notin S$ with $P(x'|y) > P(x|y)$, replacing x by x' increases the objective. Iterating yields a set of the k largest posteriors. If fewer than k elements have positive posterior, any superset up to size k that adds only zero-probability elements is equally optimal. □

A.3 Proof of Theorem 8

Since f is deterministic, $P(y|x) = \mathbf{1}\{f(x) = y\}$.

By exchangeability within $f^{-1}(y)$, there exists $\pi_y > 0$ such that $P(x) = \pi_y$ for all $x \in f^{-1}(y)$. The marginal probability is $P(y) = \sum_{x \in f^{-1}(y)} P(x) = m_f(y) \cdot \pi_y$.

For $x \in f^{-1}(y)$, the posterior is:

$$P(x|y) = \frac{P(y|x)P(x)}{P(y)} = \frac{1 \cdot \pi_y}{m_f(y) \cdot \pi_y} = \frac{1}{m_f(y)} \tag{44}$$

Thus the posterior is uniform over $f^{-1}(y)$, yielding $\max_x P(x|y) = 1/m_f(y)$. This is strictly decreasing when $m_f(y)$ increases by adding elements with positive prior (see Remark in Section 3 for the general case without exchangeability). □

A.4 Proof of Theorem 11

Let X_i be the indicator for the i-th chain being correct. By independence, $S_n = \sum_{i=1}^{n} X_i \sim \text{Binomial}(n, p)$.

For $n = 3$:

$$M_3(p) = p^3 + 3p^2(1-p) = 3p^2 - 2p^3 \tag{45}$$

We show $M_3(p) > p$ for $p \in (1/2, 1)$:

$$M_3(p) - p = p(1-2p)(p-1) > 0 \tag{46}$$

since $p > 0$, $1 - 2p < 0$ for $p > 1/2$, and $p - 1 < 0$, making the product positive.

Monotonicity in odd n follows from the identity in the main text (via Pascal's rule). By the CLT, as $n \to \infty$:

$$\frac{S_n - np}{\sqrt{np(1-p)}} \xrightarrow{d} \mathcal{N}(0, 1) \tag{47}$$

The majority is correct when $S_n > n/2$. Since $p > 1/2$, the standardized threshold $-\sqrt{n}(p - 1/2)/\sqrt{p(1-p)} \to -\infty$, so $P_{\text{majority}} \to 1$.

By Hoeffding's inequality, the error probability decays exponentially with rate $2(p-1/2)^2$. □

A.5 Proof of Theorem 14

This follows directly from the definition of strict propriety. For any true distribution $p = P(\cdot|y)$ and strictly proper scoring rule S:

$$\mathbb{E}_{x\sim p}[S(p, x)] < \mathbb{E}_{x\sim p}[S(q, x)] \quad \text{for all } q \neq p \tag{48}$$

Therefore, $q^*(\cdot|y) = P(\cdot|y)$ uniquely minimizes the expected score. □