VERIFIABLE ROBUST FEDERATED INFERENCE OF LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Organisations increasingly favour on-premise deployment of open-weight large language models (LLMs) to avoid exposing proprietary assets and private data to third-party APIs. This choice trades one risk for another: open weights can be fine-tuned, quantised, or backdoored by actors in the supply chain, and a deployed model can become a single point of failure. Federated inference, which queries a diverse ensemble of models and aggregates their outputs, is a natural mitigation, but only if the aggregation mechanism can tolerate a subset of faulty or adversarial models. In this work, we study Byzantine-robust federated inference of LLMs. Unlike prior work, we consider proper generation tasks, where outputs are not restricted to single tokens from a small label set, as in classification. We leverage two complementary signals for identifying correct outputs: a partial semantic-equivalence oracle over outputs and an external verifier induced by formal or executability properties. We combine these signals in a unified framework based on Verified Plurality with Abstention (VPA), which returns an answer only when its support exceeds a consensus threshold τ. We provide theoretical bounds on the error and abstention rates of any instantiation of the framework in terms of four factors: the faulty coincidence rate, honest-model competence, equivalence-oracle completeness, and verifier strength. In code tasks, our method achieves 0.7% error at 89.6% coverage on HumanEval+, versus 7.3% error for the best individual model, and reaches near zero error at higher τ. On MMLU, an adaptive adversary that endorses the largest incorrect cluster reduces accuracy by at most 1.5%.

1 INTRODUCTION

A hospital that cannot send patient records to a third-party API, or a company that cannot expose its codebase, has one realistic path to frontier-level LLM assistance: running open-weight models on its own hardware (Labrak et al., 2024). Because no single model dominates across tasks (Jiang et al., 2023; Tekin et al., 2024; Wang et al., 2025), the natural deployment queries several open-weight models from different developers and aggregates their answers, a setting known as *federated inference* (Dhasade et al., 2026). Open weights, however, trade a privacy risk for an integrity risk. Any actor in the supply chain can fine-tune, poison, or quantise a model so that it behaves normally on standard evaluations yet fails on targeted inputs (Wan et al., 2023; Hubinger et al., 2024; Egashira et al., 2024). Such a model is competent on clean benchmarks by construction, so it cannot be reliably screened out before deployment. The aggregator is therefore the last line of defence, and it must tolerate *Byzantine* models (Lamport et al., 1982): an unknown subset that may return arbitrary, adaptively chosen outputs.

For classification, this problem is well understood: identical labels signal agreement, and when honest models are competent, majority voting tolerates any Byzantine minority (Dhasade et al., 2026). The few Byzantine-robust schemes designed for LLMs accordingly restrict outputs to a single token or a small label set, and weight answers by quantities the models themselves supply, such as self-reported confidence or peer judgements (Zheng et al., 2025; Lee et al., 2026). Real deployments, however, ask LLMs to *generate*: programs (Chen et al., 2023; 2024), database queries (Pourreza et al., 2025; Qin et al., 2025), numerical derivations, structured actions. Here the voting signal vanishes. Two correct programs rarely coincide token-for-token (Kuhn et al., 2023), so raw frequency carries little information, while the space of wrong answers is unbounded.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Ensembling methods for generation assume honest participants, and each fails in an instructive way under attack. CodeT-style reranking (Chen et al., 2023) scores clusters by their size, which is precisely the quantity an adversary inflates. Using LLMs as judges or aggregators (Li et al., 2025; Wang et al., 2025; Pourreza et al., 2025) only relocates the robustness problem, since the judge brings systematic biases of its own (Chehbouni et al., 2025). Trained aggregators (Tekin et al., 2024) require task-specific data and single-token outputs. Most importantly, all of these methods always answer: when faulty models tie with or outvote the correct ones, the resulting error is silent.

Our key observation is that plurality voting conflates two questions that generation tasks allow us to separate: *do two outputs agree?* and *is an output correct?* Unlike classification, generation tasks often supply partial answers to both. Agreement can be decided by a semantic-equivalence oracle, for instance behavioural agreement on tests or symbolic equivalence checking. Correctness can be checked, imperfectly, by an external verifier such as test execution or a formal property. Classification is one extreme of this design space, with a perfect equivalence oracle and no verifier; code generation lies in between, with imperfect oracles on both axes.

Building on this observation, we propose *Verified Plurality with Abstention* (VPA). VPA partitions the outputs into equivalence classes, discards classes rejected by the verifier, and returns the largest remaining class only if its support reaches a threshold τ; otherwise it abstains. Abstention is what makes a guarantee possible: below the tolerance threshold, a Byzantine minority can at most prevent consensus, which turns a silent integrity failure into an observable availability failure that can be escalated to a human or a fallback. The verifier is what lifts robustness beyond the majority barrier: under worst-case collusion, no aggregator without a verifier tolerates $f \geq n/2$ Byzantine models, whereas VPA tolerates up to $n - \tau$ at a cost set by the verifier's false-accept rate.

Our contributions can be summarized as follows:

- We introduce VPA, a robust LLM federated inference mechanism that outputs a response when a verified subset of size at least τ agrees, or explicitly abstains otherwise. This abstention is critical for preventing severe failures when deploying LLMs in high-stakes domains.
- We propose a unified framework for analysing the robustness of VPA. For each generation task, we define an equivalence oracle characterized by its soundness and completeness c and a verifier characterized by its false accept rate α and false reject rates β. By leveraging these abstractions we derive upper bounds on the algorithm's error and abstention.
- Our analysis proves that increases in honest model competence and verifier strength strictly reduce both error and abstention. Meanwhile, the consensus threshold τ serves as a tunable parameter controlling the trade-off between the two.

2 Proposed Approach

2.1 Problem Statement

We consider a setting wherein n open weight LLM models are leveraged to collaboratively respond to a query x associated ground truth $Y^\star$. At each inference round, a trusted entity queries all models and gathers their responses $\{Y_1, Y_2, \cdots, Y_n\}$ to produce a final answer $\hat{Y}$. For any given generation task, we define a task instance $(\mathcal{Y}, \approx, [Y^\star])$ where $\mathcal{Y}$ is the output space and $\approx \subseteq \mathcal{Y} \times \mathcal{Y}$ is the semantic equivalence relation and $[Y^\star] \in \mathcal{Y}/\approx$ is the unique correct equivalence class, as such an output Y_i is faulty if $Y_i \notin [Y^\star]$. In this setting, an unknown subset of models $\mathcal{B} \subseteq [n]$ of size f may behave arbitrarily. The goal of this work is to ensure robustness of the inference protocol.

More concretely, for any generation task the challenge of picking a response from a pool n samples requires: (1) using some equivalence oracle to group semantically similar responses and (2) invoking a verifier to evaluate correctness. Projecting this back to LLM generated responses, the strength of theses two oracles is what dictates how challenging it is to find a suitable aggregate. That is, for any task that collapses to classification equivalence is trivial and easy to ascertain. However, no correctness oracle exists if one does not have access to the ground truth. Conversely, for tasks like code generation, beyond exact duplicates, equivalence is harder to gauge. It can however be approximated via execution: code that behaves the same is more likely performing the same functionality. Similarly, functional correctness can also be evaluated either via tests or the costly certified way using formal methods.

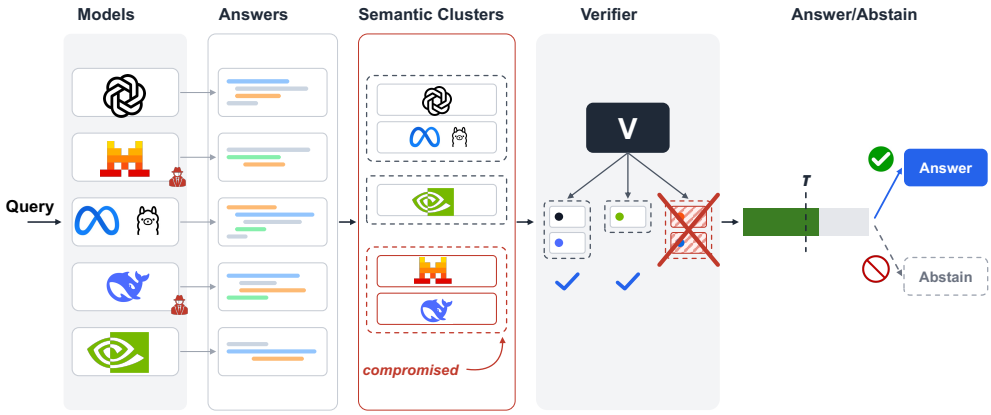

Figure 1: Overview of VPA. Responses from n models, some possibly compromised (red), are grouped into semantic clusters by an equivalence oracle E. Clusters rejected by the verifier V are discarded, and the largest remaining cluster is returned if its support reaches τ; otherwise VPA abstains.

By leveraging these ideas, to each generation task we associate an equivalence oracle and a verifier defined as follows:

Definition 2.1 (Equivalence Oracle). An *equivalence* is a procedure $\mathrm{E} : \mathcal{Y} \times \mathcal{Y} \to \{0, 1\}$ with:

- Soundness $(1 - \epsilon)$: for $Y_i \not\approx Y_j$, $\mathbb{P}[\mathrm{E}(Y_i, Y_j) = 1] \leq \epsilon$.
- Completeness (c): for $Y_i \approx Y_j$, $\mathbb{P}[\mathrm{E}(Y_i, Y_j) = 1] \geq c$.

Definition 2.2 (Verifier). A verifier is a procedure $\mathrm{V} : \mathcal{Y} \to \{0, 1\}$ with:

- False accept rate α: for $Y_i \notin [Y^\star]$, $\mathbb{P}[V(Y_i) = 1] \leq \alpha$.
- False reject rate β: for $Y_i \in [Y^\star]$, $\mathbb{P}[V(Y_i) = 0] \leq \beta$.

2.2 Verified Plurality with Abstention

In this section, we introduce Verified Plurality with Abstention (VPA), an aggregation framework that maps the n responses to either an answer or an explicit abstention, $\mathcal{A}_\tau : \mathcal{Y}^n \to \mathcal{Y} \cup \{\perp\}$. It first partitions the responses into semantic classes using E, then retains the classes that are sufficiently supported and accepted by V, and finally returns the largest retained class or abstains (Algorithm 1). E is evaluated on every pair of responses, yielding the equivalence graph G_E on vertex set $[n]$, with an edge (i, j) whenever $E(Y_i, Y_j) = 1$. The candidate classes are the connected components $C_1, \ldots, C_r$ of G_E. Soundness of E ensures that every component lies within a single semantic class. Incompleteness can therefore split a class into several components, which increases abstention, but do not merge distinct classes.

A component is retained if it reaches the consensus threshold, $|C_k| \geq \tau$, and its answer is accepted by V. Components are disjoint, so at most $\lfloor n/\tau \rfloor$ components can reach the threshold, which limits the number of verifier calls.Finally, VPA returns a representative of the largest retained component, and abstains if no component is retained.

2.3 Theoretical Analysis

In this section we derive an upper bound on $\mathbb{P}(\hat{Y} \notin [Y^\star])$ a guarantee that the output of the VPA algorithm falls within the correct equivalence class with high probability. We start by stating the assumptions followed by the main theorem and its implications for select use cases.

Assumption 2.3 (Conditional Independence and Competence)**.** Let x be an instance with ground truth answer $Y^\star$.Let $\mathcal{H}$ denote the honbest set. For all $i \in \mathcal{H}$, let Y_i denote the output of the honest model i. Conditioned on the instance, the honest outputs satisfy the following:

$$\mathbb{P}(Y_1, \ldots, Y_h \mid x) = \prod_{i \in \mathcal{H}} \mathbb{P}(Y_i \mid x)$$

Algorithm 1: Verified Plurality with Abstention

Input: Equivalence oracle E, models' responses $\{Y_1, ..., Y_n\}$, verifier V plurality τ

begin
  for $i \in [n]$ **do**
    Compute $s_{i,j} \leftarrow \mathrm{E}(Y_i, Y_j)$ for all $j \in [n]$
  Construct G_E graph with edge (i, j) iff $s_{i,j} = 1$
  Extract the connected components $\{C_1, \ldots, C_z\} \leftarrow \mathrm{ConComp}(G_E)$
  Let $\mathcal{S} \leftarrow \emptyset$
  for $k \in [z]$ **do**
    Pick a representative response $\hat{Y}_k$ from each component.
    if $\mathrm{V}(\hat{Y}_k)$ *and* $|C_k| \geq \tau$
    then
      $\mathcal{S} \leftarrow \mathcal{S} \cup \{k\}$ ▷ *Keep verified candidates*
  $k^\star = \arg\max_{k \in \mathcal{S}} |C_k|$ ▷ *Find largest verified component*
  return $\hat{\boldsymbol{Y}}_{k\star}$ *or* $\perp$ *if no unique candidate was found*

Furthermore, for some known or measurable $p \in (0, 1]$, each model's competence satisfies:

$$\mathbb{P}(Y_i \in [Y^\star] \mid x) \geq p \quad \forall i \in \mathcal{H}$$

Assumption 2.4 (γ-coincidence of faulty outputs)**.** For $S \subseteq \mathcal{H}$ with $|S| = r \geq 1$, consider

$$F_S := \{\, Y_i \notin [Y^\star] \;\; \forall i \in S\},$$
$$C_S := \{\, \exists W \in \mathcal{Y}/\!\approx,\; W \neq [Y^\star] :\; Y_i \in W \;\forall i \in S\}.$$

There exists $\gamma \in [0, 1]$ such that for every such S with $\mathbb{P}(F_S) > 0$,

$$\mathbb{P}\big(C_S \mid F_S\big) \;\leq\; \gamma^{\,r-1},$$

where the probability is over the instance x and the honest models' randomness.

Remark 2.5. γ upper-bounds the probability that two faulty outputs fall in the same wrong equivalence class. The exponent $r - 1$ encodes that, once one faulty output has fixed a wrong class, each further faulty output lands in it with probability at most γ.

Assumption 2.6 (Sound Equivalence)**.** Throughout the paper, we assume E is sound and for equivalent pairs the events $\{\mathrm{E}(Y_i, Y_j) = 1\}$ are independent across pairs each with probability $\geq c$.

Assumption 2.7 (Trusted Verifier)**.** Assume that V's randomness is independent of all models' outputs and E's randomness, and its verdicts on distinct semantic classes are independent.

Theorem 2.8 (Bounded Error & Abstention)**.** *Let assumptions 2.3-2.7 hold, let* $\hat{Y} = \mathcal{A}_\tau(Y_1, \ldots, Y_n)$ *be the output of Algorithm 1, and let* $h := n - f$ *and* $u := 1 - p$*. For every* $t \in \{1, \ldots, \tau\}$ *the error* $\delta_\mathcal{E}$ *is bounded as follows:*

$$\delta_\mathcal{E} = \mathbb{P}(\hat{Y} \notin [Y^\star]) \leq \alpha \underbrace{\left(\left\lfloor \frac{f}{t} \right\rfloor + \binom{h}{\tau - t + 1} u^{\,\tau - t + 1}\, \gamma^{\,\tau - t}\right)}_{\mathcal{E}_\tau(t)}, \tag{1}$$

Moreover, for any $\delta \in (0, 1)$ *with* $k_0 := \lceil (1 - \delta)ph \rceil \geq \tau$*, and any* $t' \in \{1, \ldots, k_0\}$ *the abstention probability of algorithm 1* $\delta_\perp$ *is bounded as follows:*

$$\delta_\perp = \mathbb{P}(\mathcal{A}_\tau(\cdot) = \perp) \leq \underbrace{e^{-\delta^2 ph/2}}_{\textit{honest mass}} + \underbrace{\Phi(k_0)}_{\textit{fragmentation}} + \underbrace{\beta}_{\textit{false reject}} + \underbrace{\alpha\, \mathcal{E}_{k_0}(t')}_{\textit{accepted tie}} \tag{2}$$

The proof of the theorem and supporting lemmas are deferred to Appendix 6.1 and 6.2 respectively.

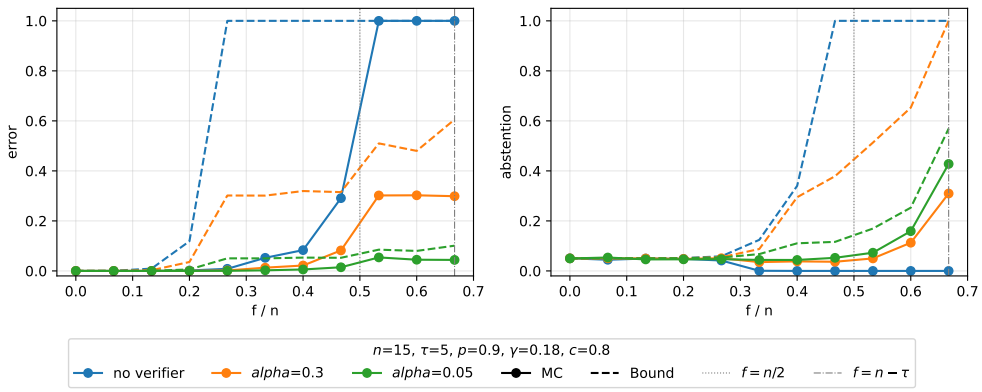

Figure 2: Theoretical error and abstention bounds vs Monte Carlo simulation (MC) under no verifier and two variants based on α

Corollary 2.9 (Worst-case adversary)**.** *Let $p = 1$, $\gamma = 1$ and $c = 1$, and assume $h \geq \tau$, i.e. $f \leq n - \tau$. The f Byzantine outputs may be arranged arbitrarily.*

1. No verifier*: ($\alpha = 1$, $\beta = 0$). If $f < h$, i.e. $f < n/2$, $\mathcal{A}_\tau$ outputs a correct answer for every arrangement. For $f \geq n/2$, no aggregator can distinguish the correct class from a wrong one: every aggregator must err or abstain in some admissible configuration.*
2. With a verifier*: For every $f \leq n - \tau$, including $f \geq n/2$,*

$$\delta_{\mathcal{E}} \leq \alpha \left\lfloor \frac{f}{\tau} \right\rfloor, \qquad \delta_{\perp} \leq \beta + \alpha \left\lfloor \frac{f}{h} \right\rfloor, \tag{3}$$

Remark 2.10. Without a verifier, this recovers the classical bound of $n/2$. Having a verifier extends tolerance to $n - \tau$ replacing the classical threshold with a guarantee that degrades with α.

Numerical validation. Figure 2 compares the bounds of Theorem 2.8 with Monte Carlo estimates for $n = 15$, $\tau = 5$, $p = 0.9$, $\gamma = 0.18$ and $c = 0.8$, as the number f of Byzantine models grows and all Byzantine outputs collude in a single wrong class. Three regimes emerge. For $f < \tau$, the Byzantine class cannot reach the threshold on its own. The error is near zero with or without a verifier, and abstention stays near 5%. For $\tau \leq f < n/2$, the Byzantine class alone meets the threshold. At $p < 1$, the correct class can fall below f, so Theorem 2.8 allows errors before $n/2$. A verifier caps them at roughly α, since only $\lfloor f/\tau \rfloor = 1$ faulty class can reach the threshold. In simulation at $f = 7$, the aggregator without a verifier errs on 29% of instances, whereas with $\alpha = 0.05$ error stays near 1% and abstention absorbs the difference. Beyond $n/2 < f \leq n - \tau$, honest models can no longer match the Byzantine class. Without a verifier the adversary then always wins, in line with the $n/2$ barrier of Corollary 2.9. With one, error saturates at α and the remaining mass goes to abstention, so VPA fails safe rather than silently. The error bound is tight in the first and third regimes and conservative only in the transition.

For the abstention bound, we evaluate the honest-mass term as the exact binomial tail $\mathbb{P}[\mathrm{Bin}(h, p) < k_0]$. This tail is the intermediate quantity in the proof of Lemma 6.2 and is tighter than its Chernoff relaxation in Theorem 2.8.

3 Empirical Evaluation

3.1 Set up

Datasets We evaluate our approach on the following benchmark datasets: We use a subset of Massive Multitask Language Understanding (MMLU) Hendrycks et al. (2021) to capture classification tasks, MedCalc-Bench Khandekar et al. (2024) a dataset of medical calculation tasks that evaluates LLM reasoning on single token generation (results are deferred to the Appendix 7) and HumanEval+ Liu et al. (2023) for coding tasks.

Models We use a diverse set of open weight models from 11 different providers. A detailed table of the models is deferred to the appendix Table 2.

Byzantine model Adaptive Byzantine are added to the response pool by simulating malicious responses either to push abstention or flip the winning component to the adversary's advantage.

3.2 Code Generation

To split code into semantic equivalent clusters, we the following equivalence oracles:

1. CodeBLEU (CB): a weighted combination of n-gram overlap s_b, keyword-weighted n-gram overlap s_{wb}, Abstract Syntax Tree (AST) match s_{AST} and Data Flow (DF) match s_{DF}. Throughout the experiment we only use the AST and DF and compute equivalence for any two pairs as follows: $\text{E}_{\text{CB}}(\text{P}_1, \text{P}_2) = \mathbb{1}\left[\text{CB}(P_1, P_2) \geq \kappa\right]$, where $\kappa = 0.9$ and $\text{CB} = \frac{1}{2}(\text{s}_{\text{AST}} + \text{s}_{\text{DF}})$.
2. LLM-generated Tests: We use a separate LLM (**deepseek-v4-flash**) to generate a subset of tests based on the task prompts. These tests are leveraged for behavioural similarity clustering of candidate solutions. In the experiments **Tests** are generated with reasoning disabled, then a second set of tests denoted throughout the experiment as **RTests** are generated with reasoning enabled to cover edge cases and improve test quality.

$$\text{E}_{\text{T}}(\text{P}_1, \text{P}_2) = \mathbb{1}\left[\forall t \in T : \text{P}_1(t) = \text{P}_2(t)\right]$$

3. CrossHair (CH): We use crosshair differential analysis (`crosshair diff-behavior`) on pairwise candidate solutions. CrossHair symbolically executes both functions and searches for an input on which their observable behaviour differs, within a fixed time budget $\Delta = 30s$.

$$E_{\text{CH}}(P_1, P_2) = \begin{cases} 1, & \text{search exhausted, no counterexample found} \\ 0, & \exists x : P_1(x) \neq P_2(x) \text{ (counterexample) or timeout} \end{cases} \tag{4}$$

i.e. $E_{\text{CH}}(P_1, P_2) = 1$ only if CrossHair terminates having exhausted its search without producing a differing input; both refutation and inconclusive runs are conservatively treated as non-equivalent.

We use number of passed tests as verification oracle.

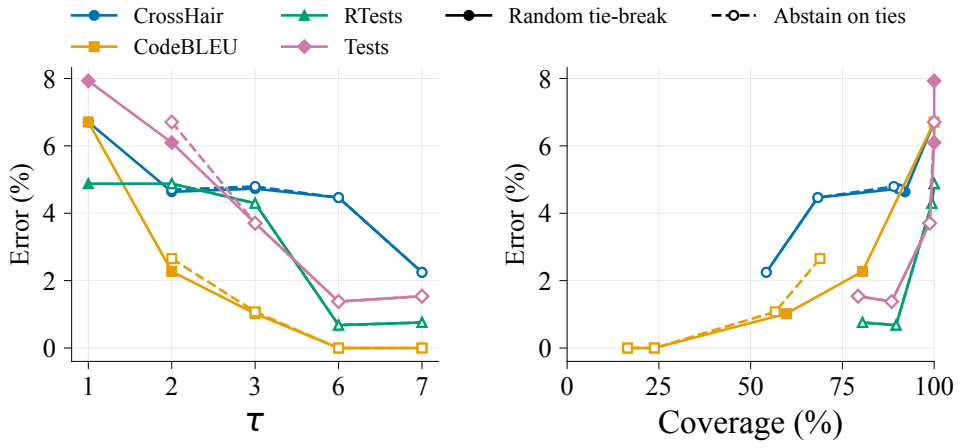

Figure 3: Comparison of varying equivalence oracles on HumanEval+ with varying plurality τ. left figure illustrates the impact of τ on error and the right figure the incurred coverage loss.

Figure 3 illustrates the algorithm's error vs coverage trade-off and how its impacted by the threshold τ. We compare the error of our method under three varying equivalence oracles: error goes down with τ for all four variants of the method. This in turn drives up abstention that results in trading coverage for lower error rates.

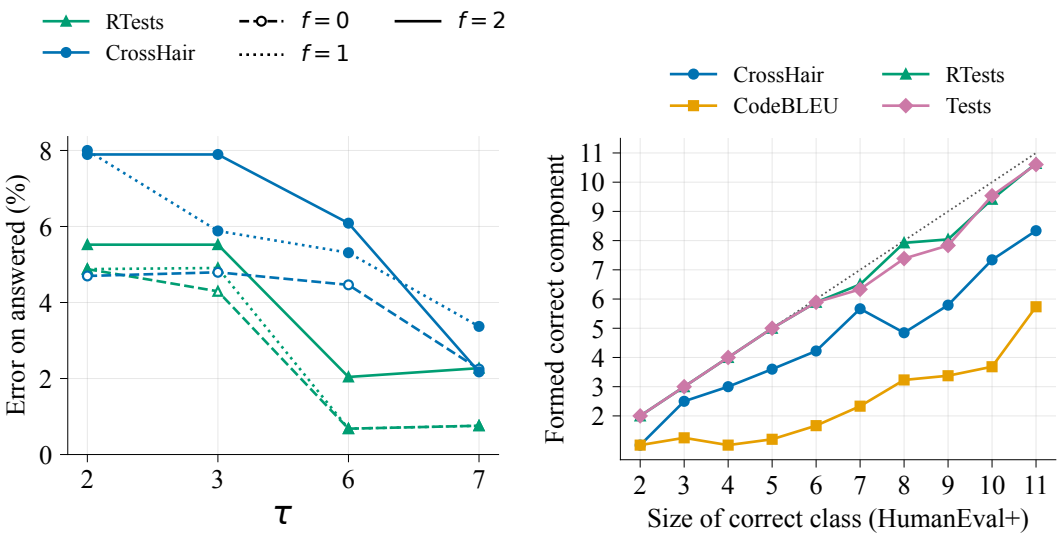

Figure 4: (left figure) HumanEval performance of two variants under an adaptive adversary that supports the second largest component. (right figure) the size of the true correct component from models that pass HumanEval+, versus the largest correct cluster built by each equivalence oracle. The dotted diagonal corresponds to an oracle that recovers the entire correct class; points below it indicate that correct solutions are fragmented.

Although, CodeBLEU yields the lowest error rate this achieved at a very low coverage: for $\tau = 2$ the error is bellow 3% with coverage already bellow 70% compared to VPA variant using RTests (Table 1) that maintain coverage above 80% with error bellow 2%.

Not all methods experience the same level of coverage degradation, this is due to the variance in completeness among the equivalence oracles. These can be loosely ranked based on their completeness as follows: $c_{tests} > c_{ch} > c_{cb}$. This ordering is apparent in right Figure 4 wherein CodeBLEU generates smaller components compared to the other equivalence oracle and to the ground truth correct component. This results in fragmention of the correct class and leads to more abstention.

In Figure 5, in the absence of an active adversary, VPA and CodeT perform comparably at full coverage. Under attack, however, CodeT degrades twice as fast with $f = 2$, because its score rewards cluster size, which the adversary targets. Because VPA uses the number of passed tests and consensus size separately and with its ability to abstain it is more robust than CodeT.

3.3 Classification Tasks

Classification tasks correspond to an extreme point where E is exact and complete, but no verifier exists, so VPA reduces to thresholded plurality voting and inherits its $n/2$ ceiling. We consider generation tasks where a single token is generated. Usually these tasks can be easily cast as classification wherein the output space is defined as $\mathcal{Y} = \{1, \cdots, K\}$.

To test this, we pick a set of diverse models $n = 7$ and compare the performance of the algorithm under varying consensus thresholds $\tau \in \{3, 4, 5, 6\}$ in two cases: (1) Faulty models due to low competence Figure 6. (2) Simulated malicious behavior that occurs due to poisoned models intentionally inducing errors Figure 7.

In Figure 6, as we go from $\tau = 3$ to $\tau = 5$ the error of the algorithm goes down by 45% as abstention goes up. This perfectly illustrates the theorem's prediction: a higher τ will explicitly trade an increased abstention rate for a lower overall error.

In Figure 7 we add an adaptive Byzantine that identifies the largest incorrect cluster and supports it. This targeted attack specifically impacts edge cases where the faulty component needs exactly one more supporter to disrupt the consensus.

Table 1: HumanEval+ comparing the use of test-based behavioral clustering as an equivalence oracle, then number of passed tests as a verification oracle. Impact of abstention on error compared to base models and error (%) of base models when forced to abstain in the same tasks as the method

	No abstention	Abstain where VPA abstains (τ)			
		2	3	6	7
nemotron-3-nano-30b-a3b	7.3	7.3	6.7	4.1	3.0
gpt-oss-20b	7.3	7.3	6.7	2.7	1.5
deepseek-v3.2-exp	7.3	7.3	6.7	3.4	3.0
gpt-oss-120b	8.5	8.5	8.0	4.1	3.8
qwen3-coder-30b-a3b-instruct	12.2	12.2	11.7	6.8	3.8
cohere_north-mini-code	16.5	16.5	16.0	10.2	4.5
glm-4.5-air	17.7	17.7	17.2	10.9	6.1
llama-3.3-70b	22.0	22.0	21.5	17.0	12.9
laguna-s-2.1	24.4	24.4	23.9	18.4	12.1
llama-3.1-8b-instruct	34.8	34.8	34.4	31.3	25.8
google_gemma-4-26b-a4b-it	54.3	54.3	54.0	50.3	45.5
Best base model	7.3	7.3	6.7	2.7	1.5
VPA (RTests + abstain on ties)	–	**4.9**	**4.3**	**0.7**	**0.8**
Coverage (%)	100.0	100.0	99.4	89.6	80.5
Tasks answered	164	164	163	147	132

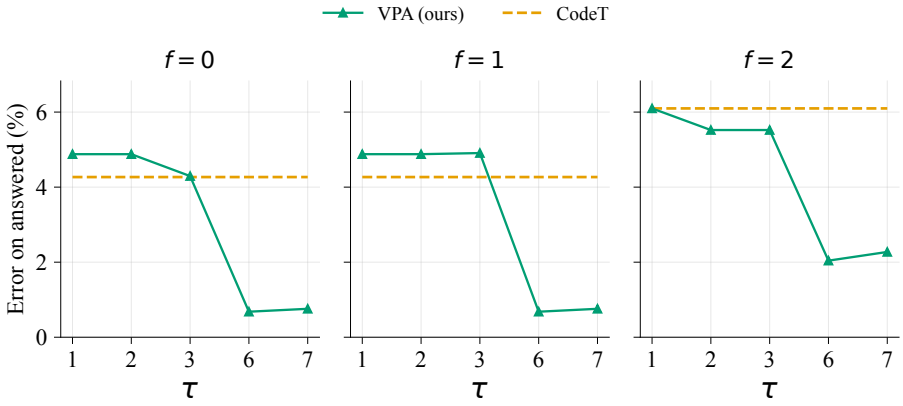

Figure 5: Comparing error of VPA variant with RTests to a variant of CodeT Chen et al. (2023) adapted to federated inference under an adaptive adversary for $f \in \{1, 2\}$ perform an adaptive attack supporting the second largest component.

This dynamic is empirically visible at higher thresholds. At ($\tau \geq 5$) the attack fails to takeover the correct consensus; for instance, the correct rate remains perfectly static between the baseline and the attack at 78.5% for $\tau = 5$ and 67.8% for $\tau = 6$. Instead, the Byzantine node breaks the abstentions, converting them directly into errors. At $\tau = 5$, the 3.6% increase in incorrect responses mirrors the exact 3.6% drop in abstentions.

4 Discussion

We propose VPA, a robust LLM federated inference framework which separates two questions that plurality voting conflates. Whether outputs agree is decided by an equivalence oracle E; whether an output is correct is decided by a verifier V. An answer is returned only when a verified class reaches support τ and we bound error and abstention for any task in terms of four quantities: the coincidence rate γ of faulty outputs, honest competence p, the completeness c of E, and the error rates (α, β) of

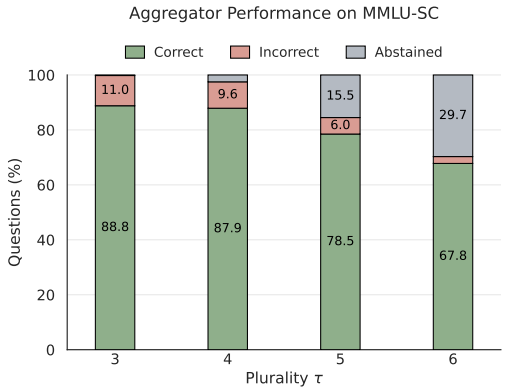

Figure 6: MMLU with varying abstention threshold τ with $n = 7$ base models

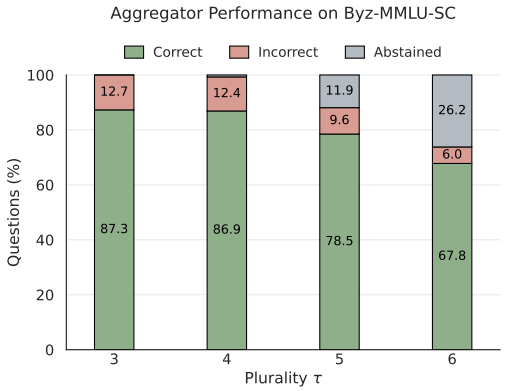

Figure 7: MMLU with varying abstention τ with $n = 8$ and an adaptive Byzantine

V. To our knowledge, this is the first Byzantine-robust aggregation scheme for LLM inference with provable error and abstention guarantees beyond classification outputs.

Importance of abstention. More fundamentally, we argue that abstention is an appropriate, and often necessary, response in uncertain settings where a wrong answer costs more than no answer, such as code shipped to production, clinical decisions, or military decision-making. VPA further shows that abstention limits what an adversary can achieve: below the tolerance threshold, a Byzantine minority can at most prevent consensus. An integrity failure thus becomes an availability failure, which is observable and can be escalated to a fallback. Interestingly, this argument does not weaken as models improve. A higher p tightens the honest-mass term and reduces abstention. The error bound, however, is driven by f, γ and α, which measure compromise rather than capability, and a backdoored model is by design competent on clean inputs.

Limitations. While VPA provides provable error and abstention guarantees for any task that supplies an equivalence oracle and a verifier, these guarantees hold under idealised assumptions that practical instantiations meet only to varying degrees.First, Theorem 2.8 assumes a sound equivalence oracle, a property that empirical oracles satisfy only approximately. Syntactic similarity, agreement on a finite test set, and time-bounded symbolic execution may each declare non-equivalent outputs equivalent. The strong empirical results obtained with these oracles nonetheless suggest that approximate soundness is often sufficient in practice. Carrying the soundness parameter ϵ of Definition 2.1 through the analysis would bring them formally within the scope of the theory.

Second, robustness has a price in computation and coverage. VPA requires n inferences and $O(n^2)$ equivalence queries. This cost can be reduced by running cheap canonicalisation or syntactic checks before expensive symbolic ones. Coverage decreases as τ grows, but the trade-off is explicit and tunable. Adapting τ to the cost of deferral in each application is a natural next step.

More broadly, the framework applies wherever outputs admit a meaningful notion of equivalence. Extending it to domains such as agentic workflows, where both sound equivalence checks and strong verifiers are available, is the most promising direction for future work.

5 Conclusion

We studied Byzantine-robust federated inference of LLMs beyond classification. VPA combines an equivalence oracle, a verifier, and a consensus threshold, and returns an answer only when a verified class gathers sufficient support. Our analysis bounds error and abstention for any instantiation. It also shows that a verifier can raise the tolerance from $n/2$ to $n-\tau$. Empirically, VPA outperforms the best individual model on code generation and matches it on medical calculation, without knowing in advance which model is best. It also resists adaptive adversaries while trading coverage for reliability in a controlled way. These results support aggregation with explicit abstention as a principled basis for deploying open-weight LLMs where a single model is an unacceptable point of failure.

AI USE STATEMENT

In this work, we used LLM-based assistants only for language polishing and grammar checking of the manuscript. They were not used to generate research ideas, derive theoretical results, design experiments, or analyse data. The authors reviewed all edited text and take full responsibility for the content of the paper.

ETHICS STATEMENT

This work aims to make LLM inference more reliable when some models in a deployment are faulty or have been compromised. No human subjects were involved in this research. All experiments use publicly released benchmarks (MMLU, HumanEval+, and MedCalc-Bench) and publicly available open-weight models, each used under its licence. MedCalc-Bench is built from publicly released, de-identified clinical text, and we did not collect or process any private patient data. The adversaries in our evaluation are simulated. They copy or support existing model outputs, and we do not release any poisoned or backdoored models. Our results on medical calculation show that abstention reduces error, but they do not mean VPA is ready for clinical use. Any deployment in a safety-critical domain would still need domain-specific validation and human oversight of abstained and answered queries.

REPRODUCIBILITY STATEMENT

We have made every effort to ensure our results can be reproduced. **Theory.** All assumptions are stated explicitly and complete proofs of Theorem 2.8, Corollary 2.9 and all supporting lemmas appear in Appendices 6.1 and 6.2 **Experiments.** Section 3 describes the benchmarks (MMLU, HumanEval+ and MedCalc-Bench) and how the adaptive Byzantine adversary is constructed. Table 2 lists every open-weight model we use, its developer, and the benchmarks on which it is evaluated. **Code and data.** Code to reproduce all experiments and figures will be released in an anonymous GitHub repository: `https://anonymous.4open.science/r/VPA-E1AC`.

REFERENCES

Khaoula Chehbouni, Mohammed Haddou, Jackie CK Cheung, and Golnoosh Farnadi. Neither valid nor reliable? investigating the use of LLMs as judges. In *The Thirty-Ninth Annual Conference on Neural Information Processing Systems Position Paper Track*, 2025. URL https://openreview.net/forum?id=yqKfMr0yvY.

Bei Chen, Fengji Zhang, Anh Nguyen, Daoguang Zan, Zeqi Lin, Jian-Guang Lou, and Weizhu Chen. Codet: Code generation with generated tests. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=ktrw68Cmu9c.

Jingchang Chen, Hongxuan Tang, Zheng Chu, Qianglong Chen, Zekun Wang, Ming Liu, and Bing Qin. Divide-and-conquer meets consensus: Unleashing the power of functions in code generation. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=cFqAANINgW.

Akash Dhasade, Sadegh Farhadkhani, Rachid Guerraoui, Nirupam Gupta, Maxime Jacovella, Anne-Marie Kermarrec, and Rafael Pinot. Robust federated inference. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=47eKYCaBIV.

Kazuki Egashira, Mark Vero, Robin Staab, Jingxuan He, and Martin Vechev. Exploiting LLM quantization. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Evan Hubinger, Carson Denison, Jesse Mu, Mike Lambert, et al. Sleeper agents: Training deceptive LLMs that persist through safety training, 2024. URL https://arxiv.org/abs/2401.05566.

Dongfu Jiang, Xiang Ren, and Bill Yuchen Lin. LLM-blender: Ensembling large language models with pairwise ranking and generative fusion. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14165–14178, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.792. URL https://aclanthology.org/2023.acl-long.792/.

Nikhil Khandekar, Qiao Jin, Guangzhi Xiong, Soren Dunn, Serina S Applebaum, Zain Anwar, Maame Sarfo-Gyamfi, Conrad W Safranek, Abid Anwar, Andrew Jiaxing Zhang, Aidan Gilson, Maxwell B Singer, Amisha D Dave, R. Andrew Taylor, Aidong Zhang, Qingyu Chen, and Zhiyong Lu. Medcalc-bench: Evaluating large language models for medical calculations. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL https://openreview.net/forum?id=VXohja0vrQ.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty quantification in language models. In *International Conference on Learning Representations*, 2023.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. BioMistral: A collection of open-source pretrained large language models for medical domains. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 5848–5864, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.348. URL https://aclanthology.org/2024.findings-acl.348/.

Leslie Lamport, Robert Shostak, and Marshall Pease. The byzantine generals problem. *ACM Trans. Program. Lang. Syst.*, 4(3):382–401, July 1982. ISSN 0164-0925. doi: 10.1145/357172.357176. URL https://doi.org/10.1145/357172.357176.

Haejoon Lee, Vincent-Daniel Yun, Hyeonho Oh, Dimitra Panagou, and Sai Praneeth Karimireddy. Robust multi-agent LLMs under byzantine faults. In *Second Workshop on Agents in the Wild: Safety, Security, and Beyond*, 2026. URL https://openreview.net/forum?id=MgHy7f7VxT.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. From generation to judgment: Opportunities and challenges of LLM-as-a-judge. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 2757–2791, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.138. URL https://aclanthology.org/2025.emnlp-main.138/.

Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and LINGMING ZHANG. Is your code generated by chatGPT really correct? rigorous evaluation of large language models for code generation. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=1qvx610Cu7.

Mohammadreza Pourreza, Hailong Li, Ruoxi Sun, Yeounoh Chung, Shayan Talaei, Gaurav Tarlok Kakkar, Yu Gan, Amin Saberi, Fatma Ozcan, and Sercan O Arik. CHASE-SQL: Multi-path reasoning and preference optimized candidate selection in text-to-SQL. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=CvGqMD5OtX.

Yang Qin, Chao Chen, Zhihang Fu, Ze Chen, Dezhong Peng, Peng Hu, and Jieping Ye. ROUTE: Robust multitask tuning and collaboration for text-to-SQL. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=BAglD6NGy0.

Selim Furkan Tekin, Fatih Ilhan, Tiansheng Huang, Sihao Hu, and Ling Liu. LLM-TOPLA: Efficient LLM ensemble by maximising diversity. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (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/.

Alexander Wan, Eric Wallace, Sheng Shen, and Dan Klein. Poisoning language models during instruction tuning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, 2023.

Junlin Wang, Jue WANG, Ben Athiwaratkun, Ce Zhang, and James Zou. Mixture-of-agents enhances large language model capabilities. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=h0ZfDIrj7T.

Lifan Zheng, Jiawei Chen, Qinghong Yin, Jingyuan Zhang, Xinyi Zeng, and Yu Tian. Rethinking the reliability of multi-agent system: A perspective from byzantine fault tolerance, 2025. URL https://arxiv.org/abs/2511.10400.

6 APPENDIX

In this section we derive an upper bound on $\mathbb{P}(\hat{Y} \notin [Y^\star])$ a guarantee that the output of VPA falls within the correct equivalence class with high probability. We start by stating and proving the main theorem.

6.1 THEORETICAL ANALYSIS

All probabilities are taken jointly over the instance x, the models' randomness, and the randomness of E and V. We go through the proof of Theorem 2.8 bellow:

Proof. (1) Consider the case where $\mathcal{A}_\tau$ does not abstain and outputs $\hat{Y}$:

$$\mathcal{A}_\tau(Y_1, \cdots, Y_n) = \hat{Y} \notin [Y^\star] \tag{5}$$

It follows that a detected faulty class D of size $|D| \geq \tau$ was accepted by the verifier V. Invoking lemma 6.5 for $m = \tau$ gives the result.
(2) Define the good event $G = G_1 \cap G_2 \cap G_3 \cap G_4$:

$$G_1 = \{K \geq k_0\} \tag{6}$$
$$G_2 = \{\text{the } K \text{ correct outputs form one detected class } C^\star\} \tag{7}$$
$$G_3 = \{V \text{ accepts } C^\star\} \tag{8}$$
$$G_4 = \{\text{no faulty detected class of size} \geq k_0 \text{ is accepted by } V\} \tag{9}$$

On the event G: $C^\star$ is accepted with

$$|C^\star| = K \geq k_0 \geq \tau \tag{10}$$

every other accepted class is faulty. Note all detected classes inside $[Y^\star]$ are merged into $C^\star$ on G_2, hence of size $< k_0 \leq |C^\star|$ by G_4.

So $C^\star$ is the unique largest accepted class, meets the threshold, and A_τ does not abstain.

Therefore $\delta_\perp \leq \mathbb{P}[G^c]$, and the union bound over the four lemmas (Lemma 6.2; Lemma 6.3 conditioned on G_1; Lemma 6.5(i); Lemma 6.5(ii) with $m = k_0$) yields equation 2. □

6.2 PROOF OF SUPPORTING LEMMAS

Call a component D of G_E *faulty* if its class (Lemma 6.1) is not $[Y^\star]$; equivalently, $Y_i \notin [Y^\star]$ for all $i \in D$.

Lemma 6.1 (G_E Component Purity)**.** *Let assumption 2.6 hold, then it follows that every connected component D of the graph G_E falls into a unique equivalence class $W \in \mathcal{Y}/\approx$*

Proof. Soundness: $\mathrm{E}(z, v) = 1 \Rightarrow z \approx v$. So every edge of G_E joins $\approx$ equivalent vertices. If z, v are in the same component, there's a path $z = w_0, \ldots, w_l = v$ and each step gives $w_{t-1} \approx w_t$, and the transitivity of $\approx$ gives $z \approx v$. Hence all vertices of a component lie in one class □

Lemma 6.2 (Honest Mass)**.** *Let $h := |\mathcal{H}|$ and let $K = \sum_{i \in \mathcal{H}} \mathbb{1}\{Y_i \in [Y^\star]\}$ denote the number of correct outputs. Under Assumption 2.3, for any $\delta \in (0, 1)$, we have:*

$$\mathbb{P}\big[K \le (1-\delta)ph\big] \le \exp\Big(-\frac{\delta^2 ph}{2}\Big).$$

Proof. Let $K = \sum_{i \in \mathcal{H}} Z_i$ with $Z_i := \mathbb{1}\{Y_i \in [Y^\star]\}$.

Invoking assumption 2.3, the Z_i are independent Bernoulli variables with means $p_i \ge p$.

By a standard coupling argument, K stochastically dominates $B \sim \mathrm{Bin}(h, p)$:

$\mathbb{P}[K \le t] \le \mathbb{P}[B \le t]$ for every t.

Applying the multiplicative Chernoff lower-tail bound yields, for any $\delta \in (0, 1)$,

$\mathbb{P}\big[K \le (1-\delta)ph\big] \le \mathbb{P}\big[B \le (1-\delta)ph\big] \le \exp\big(-\delta^2 ph/2\big)$.

□

Lemma 6.3 (Fragmentation Bound)**.** *Let K denote the number of correct outputs. Conditioned on the event $K = k \ge 2$, the probability that these k outputs do not form a single connected component is at most*

$$\varphi(k, c) := \sum_{i=1}^{\lfloor k/2 \rfloor} \binom{k}{i} (1-c)^{i(k-i)}. \tag{11}$$

Consequently, conditioned on the event that there are at least k_0 correct outputs ($K \ge k_0$), the probability of fragmentation is bounded by

$$\Phi(k_0) := \max_{k_0 \le k \le h} \varphi(k, c). \tag{12}$$

Proof. Conditioned on $K = k \ge 2$, let $I^\star = \{i \in \mathcal{H} \mid Y_i \in [Y^\star]\}$ be the set of indices of correct outputs, and let $G^\star$ be the sub-graph of G_E induced on $I^\star$. All the correct outputs are pairwise $\approx$-equivalent. By assumption 2.6, the edges of $G^\star$ are independent, with each edge $\{i, j\}$ present with probability $c_{ij} \ge c$.

Because graph disconnection is a monotonically decreasing property (adding edges cannot disconnect a graph), the condition $c_{ij} \ge c$ guarantees that the fragmentation probability of $G^\star$ is bounded above by that of a standard Erdős–Rényi graph $G(k, c)$ where all edge probabilities are exactly c. Hence,

$$\mathbb{P}\big[G^\star \text{ fragmented}\big] \le \mathbb{P}\big[G(k, c) \text{ fragmented}\big].$$

If $G(k, c)$ is disconnected, let S be a smallest connected component. There must be at least two components, so $i := |S| \le \lfloor k/2 \rfloor$, and no edge joins S to $I^\star \setminus S$. Thus, the event of disconnection is bounded by the union over all possible subsets:

$$\big\{G(k, c) \text{ disconnected}\big\} \subseteq \bigcup_{i=1}^{\lfloor k/2 \rfloor} \bigcup_{\substack{S \subset I^\star \\ |S| = i}} \big\{\text{no edge between } S \text{ and } I^\star \setminus S\big\}. \tag{13}$$

For a fixed subset S with $|S| = i$, there are $i(k-i)$ potential crossing pairs.

Since each is absent independently with probability $1-c$, the inner event has probability $(1-c)^{i(k-i)}$.

Applying a union bound by summing over the $\binom{k}{i}$ sets of size i and over all valid sizes i gives $\varphi(k, c)$.

$$\mathbb{P}\big[G^\star \text{ fragmented}\big] \le \sum_{i=1}^{\lfloor k/2 \rfloor} \binom{k}{i} (1-c)^{i(k-i)}.$$

The bound $\Phi(k_0)$ for the event $K \ge k_0$ follows immediately by taking the worst-case upper bound of $\varphi(k, c)$ over all possible valid sizes $k \in [k_0, h]$ via the law of total probability. □

Consider a subset $D \subseteq [n]$ and let $b_D := |D \cap \mathcal{B}|$ denote its number of Byzantine indices. A connected component D of G_E is considered faulty if its class (Lemma 6.1) is not $[Y^\star]$, equivalently $Y_i \notin [Y^\star]$ for all $i \in D$. For $m \ge 1$ let

$$N_m := \#\big\{\, D \text{ faulty component of } G_E : |D| \ge m \,\big\}.$$

For $m \ge 1$ and $t \in \{1, \dots, m\}$ we define:

$$\mathcal{E}_m(t) := \Big\lfloor \frac{f}{t} \Big\rfloor + \binom{h}{m-t+1} u^{\,m-t+1}\, \gamma^{\,m-t}, \qquad \mathcal{E}_m := \min_{1 \le t \le m} \mathcal{E}_m(t). \tag{14}$$

where the first term bounds the number of components with at least t Byzantine outputs, and the second bounds the expected number of r-subsets of honest outputs that are all faulty and coincide.

Lemma 6.4 (Faulty coincidence tail)**.** *Let Assumptions 2.3, 2.4 and 2.6 hold. For every $m \ge 1$ and $t \in \{1, \dots, m\}$*

$$\mathbb{P}\big[\exists \textit{ faulty component } D :\ |D| \ge m\big] \le \mathcal{E}_m(t).$$

Proof. The event that some faulty component has size at least m is exactly $\{N_m \ge 1\}$. Since N_m is a non-negative integer, $\mathbb{1}\{N_m \ge 1\} \le N_m$ pointwise, so

$$\mathbb{P}\big[\exists \text{ faulty component } D :\ |D| \ge m\big] = \mathbb{P}[N_m \ge 1] \le \mathbb{E}[N_m].$$

It therefore suffices to show $\mathbb{E}[N_m] \le \mathcal{E}_m(t)$ for every fixed $t \in \{1, \dots, m\}$.

Fix $t \in \{1, \dots, m\}$ and split the faulty components of size $\ge m$ into Case(1): $b_D \ge t$ and Case(2): $b_D \le t-1$ so that $N_m = N_m^{(1)} + N_m^{(2)}$.

Case(1)

Because components are disjoint, we have $tN_m^{(1)} \le \sum_D b_D \le f$.

Hence, $N_m^{(1)} \le \lfloor f/t \rfloor$ deterministically, with no assumption on the Byzantine outputs.

Case (2)

Each faulty component D satisfies $|D \cap \mathcal{H}| \ge m - t + 1 =: r$. By Lemma 6.1, D lies in a single class $W \ne [Y^\star]$, so every $S_D \subseteq D \cap \mathcal{H}$ with $|S_D| = r$ satisfies $F_{S_D} \cap C_{S_D}$. Because components are disjoint the S_D are distinct, hence

$$\mathbb{E}\big[N_m^{(2)}\big] \le \sum_{\substack{S \subseteq \mathcal{H} \\ |S| = r}} \mathbb{P}[F_S \cap C_S] \ \le\ \binom{h}{r} u^r\, \gamma^{r-1}.$$

Invoking (Assumption 2.4) we have: $\mathbb{P}[C_S \mid F_S] \le \gamma^{r-1}$ and $\mathbb{P}[F_S] \le u^r$, which follows from conditional independence and per-instance competence (Assumption 2.3).

Finally, adding the two cases and minimizing over t gives:

$$\mathbb{E}[N_m] \le \mathcal{E}_m$$

□

Lemma 6.5 (Verifier Error)**.** *Let Assumptions 2.3–2.7 hold. Then:*

(i) Each detected class contained in $[Y^\star]$ is rejected with probability at most β.
(ii) For every $m \geq 1$ and $t \in \{1, \ldots, m\}$,

$$\mathbb{P}\big[\exists \textit{ faulty detected class of size} \geq m \textit{ accepted by V}\big] \leq \alpha\, \mathcal{E}_m(t).$$

Proof. (i) By Lemma 6.1, the representative lies in $[Y^\star]$. Since V's randomness is independent of the outputs and E (Assumption 2.7), Definition 2.2 applies directly.

(ii) Let M_m count the accepted faulty classes of size at least m; the event is $\{M_m \geq 1\}$, so its probability is at most $\mathbb{E}[M_m]$. Given the outputs and E, each such class has its representative outside $[Y^\star]$ and is accepted with probability at most α. By linearity, $\mathbb{E}[M_m] \leq \alpha\, \mathbb{E}[N_m] \leq \alpha\, \mathcal{E}_m(t)$ by Lemma 6.4. □

Remark 6.6 (Classification)*.*

- In classification, E is exact label matching, hence it is both sound and complete ($c = 1$, so $\Phi \equiv 0$), and no verifier is available ($\alpha = 1$, $\beta = 0$). $\mathcal{A}_\tau$ reduces to plurality voting with threshold τ, and Theorem 2.8 gives, for every $t \leq \tau$ and $t' \leq k_0$,

$$\delta_{\mathcal{E}} \leq \mathcal{E}_\tau(t), \qquad \delta_\perp \leq e^{-\delta^2 p h/2} + \mathcal{E}_{k_0}(t').$$

  If $f < \tau$ the adversary will need $\tau - f$ honest models to fail and coincide. If $f \geq \tau$ and without a verifier the adversary alone can reach the threshold, protection then comes only from the honest majority (Corollary 2.9).
- As $p \to 1$, $u \to 0$ and $\delta_{\mathcal{E}} \leq \lfloor f/\tau \rfloor$: the guarantees reduce to comparing the number of Byzantine models f with the thresholds τ and k_0.

Corollary 6.7 (Worst-case adversary)**.** *Let $p = 1$, $\gamma = 1$ and $c = 1$, and assume $h \geq \tau$, i.e. $f \leq n - \tau$. The f Byzantine outputs may be arranged arbitrarily.*

1. No verifier*: ($\alpha = 1$, $\beta = 0$). If $f < h$, i.e. $f < n/2$, $\mathcal{A}_\tau$ outputs a correct answer for every arrangement. For $f \geq n/2$, no aggregator can distinguish the correct class from a wrong one: every aggregator must err or abstain in some admissible configuration.*
2. With a verifier*: For every $f \leq n - \tau$, including $f \geq n/2$,*

$$\delta_{\mathcal{E}} \leq \alpha \Big\lfloor \frac{f}{\tau} \Big\rfloor, \qquad \delta_\perp \leq \beta + \alpha \Big\lfloor \frac{f}{h} \Big\rfloor, \tag{15}$$

   and if $f < h$, additionally $\delta_{\mathcal{E}} \leq \beta$.

The verifier thus moves the tolerance from $n/2$ to $n - \tau$, at a cost of $\alpha\lfloor f/\tau \rfloor$. A wrong class is kept only if it both passes the verifier and reaches the threshold τ, so at most $\lfloor f/\tau \rfloor$ adversarial classes are eligible, each passing with probability at most α. In particular, $\delta_{\mathcal{E}} = 0$ if $f < \tau$, the cost is α for $\tau \leq f < 2\tau$, and splitting into several classes helps the adversary only when $f \geq 2\tau$.

Proof. With $p = 1$ and $c = 1$, the h honest outputs are correct and, by soundness of E, form a single component $C^\star$ with $|C^\star| = h \geq \tau$. Every faulty component therefore contains only Byzantine outputs, so at most $\lfloor f/\tau \rfloor$ of them reach size τ and at most $\lfloor f/h \rfloor$ reach size h.

(1) Without a verifier, $\mathcal{A}_\tau$ returns the largest component, which is $C^\star$ whenever $f < h$. For $f \geq h$, the adversary can form a wrong class of size h, indistinguishable from an instance with the roles exchanged, so every aggregator errs or abstains in one of them.

(2) Each faulty component is accepted with probability at most α, independently of how the components form (Assumption 2.7). An error needs one of the at most $\lfloor f/\tau \rfloor$ faulty components of size at least τ to be accepted, which is bounded by $\alpha\lfloor f/\tau \rfloor$. If $f < h$, it also needs $C^\star$ to be rejected, with probability at most β. Abstention needs $C^\star$ rejected with probability at most β or a faulty component of size at least h accepted ($\alpha\lfloor f/h \rfloor$). □

Remark 6.8. Without a verifier, this recovers the classical bound of $n/2$. Having a verifier extends tolerance to $n - \tau$ replacing the classical threshold with a guarantee that degrades with α.

Provider	Model	MMLU	MedCalc	HumanEval+
Meta	Llama 3.1 8B	✓		✓
	Llama 3.3 70B	✓		✓
	Llama 4 Scout		✓	
OpenAI	GPT-OSS 20B	✓	✓	✓
	GPT-OSS 120B	✓	✓	✓
Google	Gemma 4 26B-A4B (IT)	✓		✓
NVIDIA	Nemotron 3 Nano		✓	✓
	Nemotron 3 Super	✓		
Poolside	Laguna S 2.1	✓	✓	✓
Microsoft	Phi-4		✓	
Alibaba (Qwen)	Qwen3 8B		✓	
	Qwen3-Coder 30B-A3B Instruct			✓
Mistral AI	Ministral 8B		✓	
Cohere	North Mini Code			✓
DeepSeek	DeepSeek-V3.2-Exp		✓	✓
Zhipu AI	GLM-4.5-Air			✓

Table 2: Open-weight models evaluated in this work, grouped by developer: 16 models from 11 developers. MMLU uses 7 models, MedCalc-Bench 9 and HumanEval 11

7 Additional Empirical Results

MedCalc Bench MedCalc-Bench answers are numeric. Integer-valued tasks (rule-based scores) are graded by exact match and decimal-valued tasks (equation-based) are graded correct if the answer lies in the benchmark's admissible interval $[\ell, u]$, which spans $\pm 5\%$ of the reference value. The equivalence oracle mirrors this rule:

$$E_{\mathrm{MC}}(Y_i, Y_j) = \begin{cases} \mathbb{1}\big[\,|Y_i - Y_j| \le \eta \max(|Y_i|, |Y_j|)\,\big], & \text{decimal outputs, } \eta = 0.05, \\ \mathbb{1}\big[\,Y_i = Y_j\,\big], & \text{integer outputs.} \end{cases}$$

Candidate classes are the connected components of G_E. The answer returned is the member of the winning component closest to the component median. Ties between largest components are broken uniformly at random with a fixed seed. No verifier is available ($\alpha = 1$, $\beta = 0$), so on MedCalc VPA reduces to threshold plurality voting.

For integer outputs, E_{MC} is sound and complete. For decimal outputs it is only approximately sound. The tolerance relation is not transitive, so a chain of pairwise-close answers can form a component wider than the admissible interval. Two answers within 5% of each other can also fall on opposite sides of an interval boundary. Returning the median member limits the effect of such chains.

Models Figure 9 and Table 3 report results on MedCalc-Bench, with $n = 9$ models, no verifier, and abstention on ties. Every model's error drops on the tasks VPA answers (e.g., gpt-oss-20b from 22.9% to 2.3% at $\tau = 6$). The pool has dominant weak models: five of the nine models exceed 40% error. Error on answered tasks decreases with τ and reaches zero for $\tau \ge 6$. At every τ, VPA matches the best base model evaluated on the same tasks. That model, however, can only be identified after the fact, using ground-truth labels. VPA reaches the same error without knowing which model is strongest and without relying on any single model.

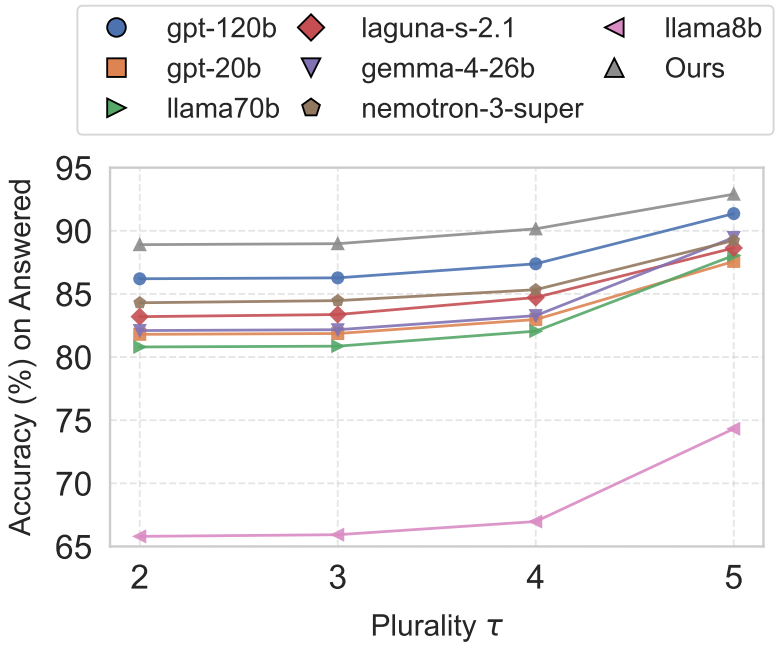

Figure 8: Accuracy vs τ on a subset of MMLU comparing VPA to $n = 7$ base models

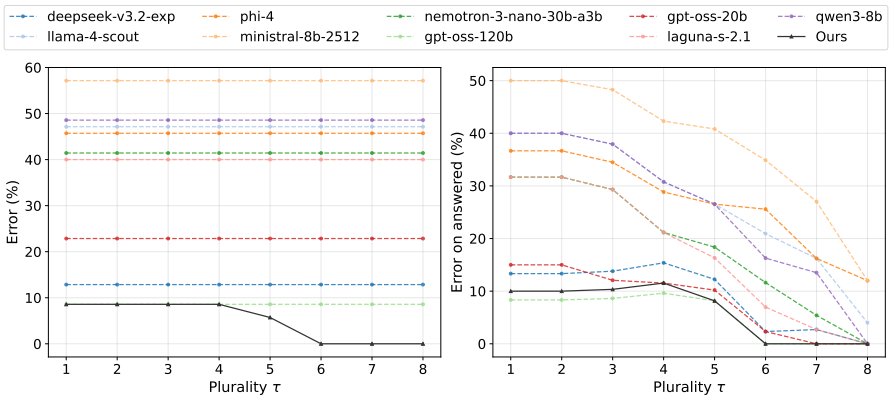

Figure 9: Error (%) on a subset of MedCalc-Bench on answered tasks. Error (%) on answered with every model evaluated on the tasks VPA answers

Table 3: Performance of VPA on MedCalc Bench compared to $n = 9$ base models under varying threshold τ

	No abstention	Abstain where VPA abstains (τ)			
		5	6	7	8
gpt-oss-120b	**8.6**	**8.2**	**0.0**	**0.0**	**0.0**
deepseek-v3.2-exp	12.9	12.2	2.3	2.7	**0.0**
gpt-oss-20b	22.9	10.2	2.3	**0.0**	**0.0**
laguna-s-2.1	40.0	16.3	7.0	2.7	**0.0**
nemotron-3-nano-30b-a3b	41.4	18.4	11.6	5.4	**0.0**
phi-4	45.7	26.5	25.6	16.2	12.0
llama-4-scout	47.1	26.5	20.9	16.2	4.0
qwen3-8b	48.6	26.5	16.3	13.5	**0.0**
ministral-8b-2512	57.1	40.8	34.9	27.0	12.0
Best base model	8.6	8.2	0.0	0.0	0.0
Ours	–	8.2	0.0	0.0	0.0
Coverage (%)	100.0	70.0	61.4	52.9	35.7

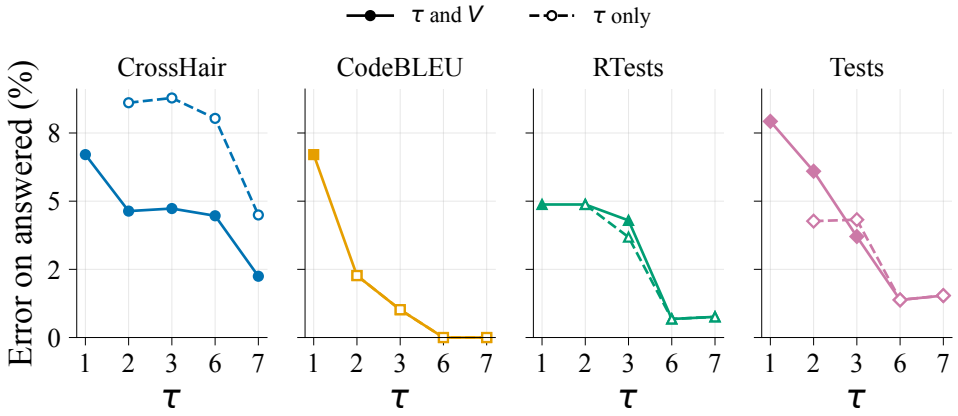

Figure 10: Comparison of performance of the method with varying equivalence oracles on HumanEval+ when paired with (τ and V) and without a verification step (τ only)

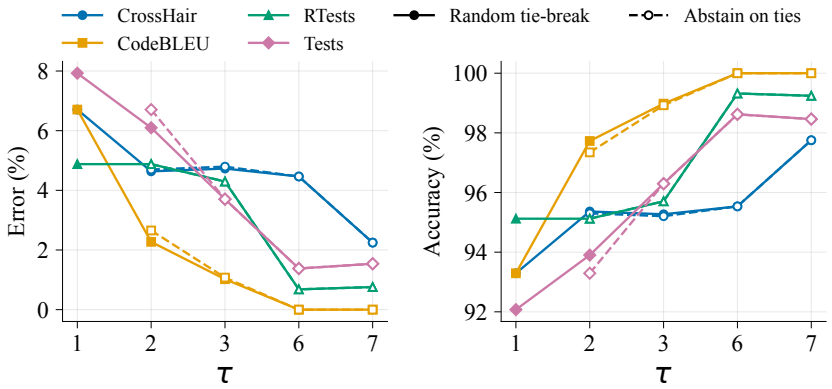

Figure 11: Comparison of varying equivalence oracles on HumanEval+ accuracy and error with varying plurality τ with and without verifier

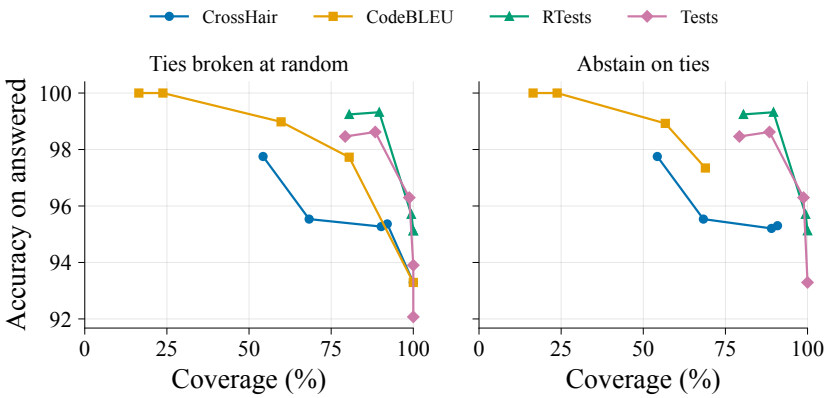

Figure 12: Comparison of varying equivalence oracles on HumanEval+ accuracy vs coverage trade-off (figure on the left) braking ties randomly and (figure on the right) our standard VPA abstaining on ties