From Answers to Hypotheses

Internal Consensus and Its Limits in Large Language Models

Victor Lavrenko
victor@peacetech.vc

Abstract

Large language models (LLMs) now reach near-expert performance on medical QA and diagnostic benchmarks, yet this progress does not reliably translate into safe clinical decision support. We argue that a core limitation is not missing knowledge, but a failure to surface and reason over the *distribution of internal hypotheses* expressed across alternative decoding trajectories. Using controlled multi-branch sampling on a medical QA benchmark, we show that (i) majority-vote self-consistency does not yield statistically significant accuracy gains over greedy decoding in our setting; (ii) the correct answer is often present among sampled hypotheses even when the final prediction is wrong; and (iii) strong internal consensus can still coincide with incorrect answers. We formalize these effects with statistical tests and characterize an alternative failure mode—*errors of hypothesis space*—where models converge confidently on a systematically wrong explanation, motivating hypothesis-centric evaluation analogous to differential diagnosis.

1 Introduction

Large language models (LLMs) have recently achieved impressive results on medical examinations and diagnostic benchmarks, in some cases rivaling or exceeding average physician performance on structured question answering and exam-style tasks. This progress has fueled optimism about near-term clinical deployment, with several systems now approaching conversational history-taking, differential diagnosis, and longitudinal reasoning. However, randomized studies and post-deployment analyses increasingly suggest that benchmark gains do not reliably translate into improved clinical decision-making and may introduce new risks when model outputs are treated as definitive answers.

A dominant explanation for this gap attributes failures to hallucinations and overconfidence. While these phenomena are real, prior work paints a more nuanced picture of model uncertainty. **kadavath2022know** show that language models often *know when they are likely to be wrong*, in the sense that internal signals correlate with error, but this knowledge is not reliably exposed through standard decoding. Similarly, **wang2022selfconsistency** demonstrate that sampling multiple reasoning traces and aggregating them via self-consistency can improve accuracy on some reasoning benchmarks, suggesting that alternative hypotheses already exist within the model's latent reasoning process.

Recent clinical and near-clinical evaluations further underscore the limits of treating improved benchmark accuracy as a proxy for dependable decision support. In a randomized controlled study, access to an LLM did not significantly improve physicians' diagnostic reasoning compared to conventional resources (**goh2024jama**), despite strong standalone model performance. At the same time, frontier systems such as AMIE illustrate rapid progress toward conversational diagnostic assistance and differential diagnosis (**tu2025amie**; **mcduff2025naturemed**), amplifying the importance of understanding when internally coherent reasoning trajectories nonetheless converge on incorrect conclusions.

Taken together, these findings suggest that uncertainty and alternative explanations are not absent from modern LLMs; rather, they are often *latent*. The open question is whether these internal structures are sufficiently reliable, diverse, and accessible to support high-stakes decision-making. In medicine,

uncertainty is rarely resolved by selecting a single most likely hypothesis; instead, clinicians reason over *sets* of competing explanations, iteratively revising them as new evidence emerges. This work adopts that perspective and studies LLM reasoning as a *distribution over hypotheses*, rather than as a single chain of thought or a single decoded answer.

Operationally, we observe this distribution through repeated sampling of the same prompt under fixed conditions, yielding an empirical distribution over final answers and associated explanations. We do not assume access to internal symbolic representations or latent states; rather, by "hypotheses" we refer to the set of discrete answer-level explanations implicitly instantiated by independent decoding trajectories. This framing allows us to analyze internal agreement, diversity, and failure modes without introducing new architectures or training procedures.

Finally, we emphasize that this work is diagnostic rather than prescriptive. We do not propose a new aggregation rule, verification strategy, or decoding algorithm. Instead, our goal is to characterize when and why aggregation and internal agreement fail, even under idealized sampling conditions, and to identify a distinct failure mode—*errors of hypothesis space*—in which models converge confidently on a systematically wrong explanation. Understanding these failures is a necessary step toward evaluation paradigms that more closely resemble differential diagnosis in clinical practice.

2 Experimental Setup

For each question, we generate:

- a greedy prediction using deterministic decoding; and
- an ensemble of $N = 10$ independently sampled reasoning paths.

Let $a_{i,j}$ denote the final answer of branch j for question i. From the empirical distribution $P_i(a)$, we compute:

- leader answer: $\arg\max_a P_i(a)$;
- maximum agreement fraction:
$$\text{max\_frac}_i = \max_a P_i(a);$$
- Top-k coverage:
$$\text{Top-}k = \mathbb{1}\{y_i \in \text{Top-}k(P_i)\}.$$

All results are computed over 400 questions using a fixed prompt and model configuration.

3 Hypothesis H1: Aggregation Improves Accuracy

H1. Majority-vote aggregation improves accuracy relative to greedy decoding.

Let $\hat{y}_i^{(g)}$ denote the greedy prediction and $\hat{y}_i^{(m)}$ the majority prediction. Accuracy is defined as:

$$\text{Acc} = \frac{1}{N} \sum_i \mathbb{1}[\hat{y}_i = y_i].$$

Empirically:

- Greedy accuracy: 65.75%
- Majority accuracy: 66.75%

A two-sided binomial test with null hypothesis $H_0 : \pi = 0.6575$, where π denotes the true success probability equal to the greedy decoding accuracy, yields a p-value of approximately 0.63. We therefore fail to reject H_0, indicating that the observed difference between greedy and majority-vote accuracy is not statistically significant in this setting.

Interpretation. These results provide no statistical evidence in support of H1 in this experimental regime, suggesting that majority-vote aggregation does not reliably improve accuracy over greedy decoding under the tested conditions. While self-consistency can improve performance in some regimes (**wang2022selfconsistency**), our results show that such gains are not robust in this setting. Aggregation alone is therefore insufficient as a general reliability mechanism.

4 Hypothesis H2: Correct Answers Appear Among Alternatives

H2. When the final prediction is incorrect, the correct answer is often present among alternative hypotheses.

We measure Top-2 coverage:

$$\text{Top-2 coverage} = \frac{1}{N} \sum_i \mathbb{1}\{y_i \in \text{Top-2}(P_i)\}.$$

Observed Top-2 coverage is 80.5%, compared to a greedy accuracy of 65.75%, corresponding to an absolute improvement of 14.75 percentage points. Using a binomial model with null hypothesis $H_0 : \pi = 0.6575$ (where π is the baseline success probability equal to the greedy accuracy), this difference is highly statistically significant (p-value $\ll 10^{-6}$).

Relation to prior work. This result is consistent with **kadavath2022know**, who show that models frequently possess internal signals of uncertainty. Our findings extend this by demonstrating that uncertainty manifests as *explicit alternative hypotheses*, not merely as reduced confidence.

5 Hypothesis H3: Internal Consensus Implies Correctness

H3. Strong internal agreement implies correctness.

We analyze unanimous cases where $\text{max\_frac}_i = 1.0$. Out of 400 questions:

- Unanimous cases: 151
- Unanimous accuracy: 86.8%

We test the null hypothesis $H_0 : \pi \geq 0.95$, where π denotes the true accuracy of unanimous predictions. A one-sided binomial test yields a p-value below 0.01, allowing us to reject this hypothesis. Thus, even strong internal consensus does not guarantee near-perfect reliability.

Key insight. Internal consensus reflects *coherence*, not truth. Models can converge confidently on incorrect explanations, producing a dangerous illusion of reliability.

6 Distributional Analysis of Consensus

Figure 1 illustrates how accuracy varies with internal consensus. Accuracy increases monotonically with agreement but saturates well below perfect reliability. Notably, near-unanimous cases ($\text{max\_frac} \geq 0.9$) still exhibit error rates above 15%.

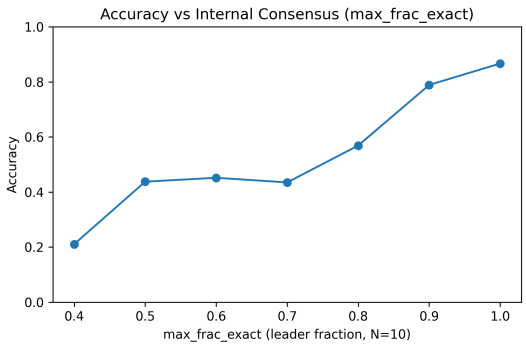

Figure 1: Accuracy as a function of internal consensus (max_frac). Higher branch agreement correlates with higher accuracy, but even near-unanimous predictions exhibit a non-zero error rate.

This behavior is inconsistent with the assumption that confidence or agreement can serve as a sufficient decision criterion.

7 An Alternative Explanation: Errors of Hypothesis Space

The failure to find evidence supporting H3 motivates an alternative explanation for these errors:

H3′. Some errors arise because the correct hypothesis is poorly represented or absent from the model's hypothesis space.

Under this view, unanimous but incorrect predictions correspond to cases where the model's hypothesis space is locally coherent but globally misaligned with the task. Such errors are not reducible to stochastic selection noise and therefore cannot be remedied by increased sampling or voting alone.

In such cases, increased sampling or stronger aggregation cannot recover the correct answer. This reframes aggregation as a *diagnostic probe* rather than a solution.

Taken together, our results suggest three qualitatively distinct failure modes in LLM reasoning:

1. Selection errors, where the correct hypothesis is present among alternatives but not selected as the final answer.
2. Calibration or confidence errors, where internal agreement or confidence is misaligned with correctness.
3. Hypothesis space errors, where the model converges on an internally coherent but systematically incorrect explanation.

Importantly, only the first class is addressable through improved selection or aggregation, while the latter two require changes to training objectives or hypothesis representation.

8 Relation to Prior Work

- **Self-consistency. wang2022selfconsistency** show that self-consistency can improve accuracy; we show its limits and failure modes in a medical QA setting.
- **Internal uncertainty. kadavath2022know** demonstrate internal uncertainty awareness; we show how this uncertainty appears as alternative hypotheses at the answer level.

- **Iterative refinement. madaan2023selfrefine** and STaR (**zelikman2023star**) focus on iterative correction and training; our work focuses on analysis rather than optimization.
- **Selective prediction / abstention.** Related questions arise in selective prediction and abstention for medical QA (**machcha2025abstention**), where the objective is to know when *not* to answer.

9 Implications and Future Work

The analysis presented in this work suggests that improving LLM performance in high-stakes medical settings is not primarily a matter of producing a single more accurate answer, but of reasoning over structured sets of competing hypotheses. In particular, the presence of meaningful top-2 coverage opens a path toward correcting errors, but only under carefully constrained decision regimes.

9.1 Limits of naive top-2 flipping

A natural first idea is to flip predictions from the most frequent answer (top-1) to the second-most frequent answer (top-2) whenever uncertainty is detected. However, our results show that such naive flipping is generally unsafe. In regimes with strong consensus, correct top-1 predictions vastly outnumber correct top-2 predictions. As a result, even a small rate of incorrect flips would outweigh any potential benefit. This imbalance is statistically significant in our data (binomial test, $p < 10^{-6}$ for high-consensus regimes), demonstrating that indiscriminate flipping would almost certainly degrade overall accuracy.

These findings indicate that top-2 coverage alone is insufficient: the feasibility of flipping depends critically on the relative frequencies of correct top-1 and top-2 hypotheses within a given uncertainty regime.

9.2 Flip feasibility and error budgets across uncertainty regimes

To make this trade-off explicit, we analyzed the feasibility of selective top-2 flipping across vote-consensus regimes defined by top-1 vote counts and top-1/top-2 gaps. For each regime, we quantified two key quantities (Figure 2).

First, we computed the maximum achievable overall accuracy under an idealized oracle that flips all cases where the second-most frequent hypothesis is correct (top-2 = gold) and never introduces new errors. This represents a strict upper bound, since errors outside top-2 coverage cannot be corrected by flipping.

Second, we measured the required suppression of false flips, defined as the ratio between cases where the top-1 hypothesis is correct and cases where the top-2 hypothesis is correct within the regime. This ratio directly reflects how accurate a flip decision policy must be in order to avoid degrading performance.

The analysis reveals a sharp, regime-dependent trade-off. Broad regimes that include high-consensus predictions offer larger theoretical gains but require extremely high flip precision, making them practically infeasible. In contrast, more conservative regimes focused on high-uncertainty cases—where top-1 and top-2 receive comparable support—exhibit substantially more favorable error budgets. In these regimes, even moderately accurate flip policies could yield meaningful improvements while incurring limited risk.

While the available evaluation set limits statistical power for fine-grained comparisons between neighboring regimes, the qualitative pattern is robust. In particular, excluding high-consensus predictions (e.g., regimes with very high top-1 vote counts) leads to a statistically significant reduction in the precision required of a flip algorithm, compared to regimes that include such cases.

We observe that moving the upper bound on top-1 vote counts from high-consensus regimes to uncertainty-focused regimes yields a more than twofold reduction in required false-flip suppression (from ≈ 2.4 to ≈ 1.0), a difference that is statistically significant ($p < 10^{-6}$).

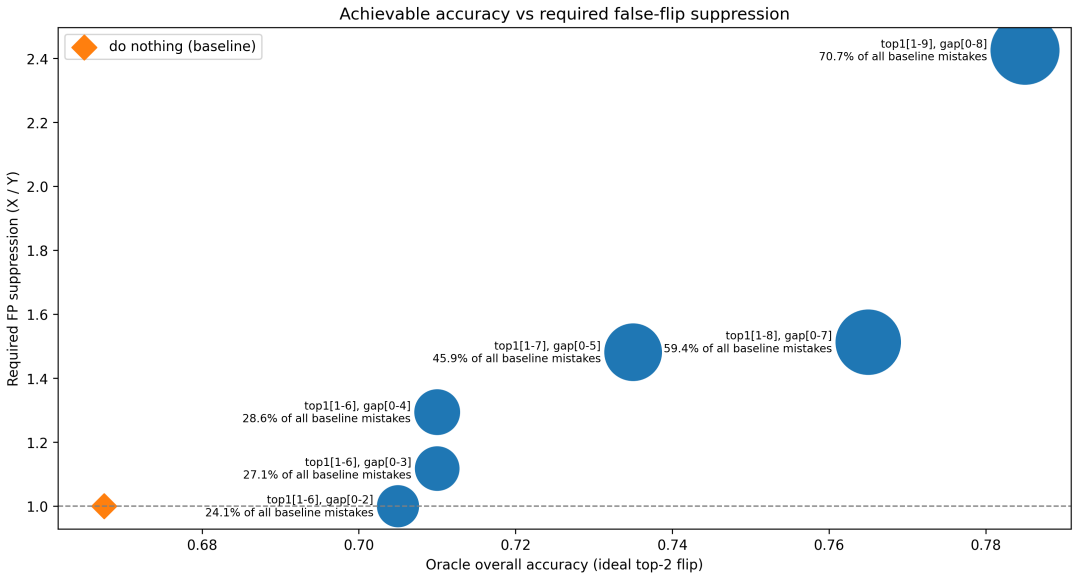

Figure 2: Feasibility of selective top-2 flipping across vote-consensus regimes. The x-axis shows the maximum overall accuracy achievable by an ideal oracle that corrects all top-2=gold cases within a regime. The y-axis indicates the required suppression of false flips (top-1 correct vs. top-2 correct). Bubble size reflects the number of baseline errors covered by the regime.

9.3 Implications for future flip algorithms

These findings suggest a staged and risk-aware strategy for future work. Rather than attempting to correct all errors, flip algorithms should initially target regimes with high uncertainty and favorable error budgets, where the cost of a mistaken flip is relatively low and the likelihood of successful correction is higher. As decision accuracy improves, such algorithms could progressively expand into more challenging regimes.

Importantly, this perspective reframes flipping not as a binary operation applied globally, but as a selective intervention governed by explicit uncertainty thresholds and error budgets. Even conservative strategies restricted to high-uncertainty regimes can already produce substantial gains, without requiring near-perfect flip decisions.

Future research should therefore focus on:

- explicit modeling of hypothesis sets and vote distributions rather than single predictions;
- learning flip decision policies that operate within predefined error budgets;
- training objectives that penalize confident convergence on incorrect hypotheses; and
- interfaces that expose structured uncertainty and alternative hypotheses to clinicians.

Taken together, these results highlight that safe and effective deployment of LLMs in medicine depends not only on accuracy, but on principled control of when and how model uncertainty is resolved.

Reproducibility and Code Availability

All experiments reported in this paper are fully reproducible. A public, versioned snapshot of the complete codebase is available at `https://github.com/victorlavrenko/rofa/tree/paper/from-answers-to-hypotheses-v1`.

Reproduction is supported by two dedicated notebooks:

- Reproduction notebook (analysis only): `20_paper_reproduce.ipynb` [Colab ↗]

- Generation notebook (optional, stochastic): `10_colab_generate.ipynb` [Colab ↗]